4th International Conference on
Development and Learning and on Epigenetic Robotics
October 13-16, 2014. Palazzo Ducale, Genoa, Italy

The Influence of Social Display in Competitive Multiagent Learning

Pedro Sequeira, Francisco S. Melo, and Ana Paiva
INESC-ID and Instituto Superior Técnico, Universidade de Lisboa
Av. Prof. Dr. Cavaco Silva
2744-016 Porto Salvo, Portugal
pedro.sequeira@gaips.inesc-id.pt, fmelo@inesc-id.pt, ana.paiva@inesc-id.pt

***Abstract*—In this paper we analyze the impact of simple social signaling mechanisms in the performance of learning agents within competitive multiagent settings. In our framework, self-interested reinforcement learning agents interact and compete with each other for limited resources. The agents can exchange social signals that influence the total amount of reward received throughout time. In a series of experiments, we vary the amount resources available in the environment, the frequency of interactions and the importance each agent in the population gives to the social displays of others. We measure the combined performance of the population according to distinct selection paradigms based on the individual performances of each agent. The results of our study show that by focusing on the social displays of others, agents learn to collectively coordinate their feeding behavior by trading-off immediate benefit for long-term social welfare. Also, given populations where the impact of the social signal on the reward differs, the individuals with the highest fitness appear in the most socially-aware populations. The presence of social signaling gives also origin to more social inequalities where the more fit agents benefit from their higher status being appreciated by other members of the population.**

I. Introduction

Social signaling exists in nature as a simple communication mechanism between individuals [1]. Signals allow for coordination behavior by informing action opportunities to others, allowing groups of agents to perform better as a whole than if they acted on their own (*e.g.*, cooperative hunting, gathering, breeding, etc.). In a *sender-receiver* perspective [2], signals carry information about the sender's observed state such that it may change the receiver's actions towards dealing with such state. Repeated interactions between signaling individuals make each adjust its own strategy towards maximizing its payoff given some fitness function [1]. In this paper we study the impact of social signaling mechanisms in the combined performance of interacting learning agents competing in an environment with limited resources.

We model agents exchanging simple social signals that inform others about their current internal state. In a sense, they function as a form of social display, possibly influencing the actions of others over time.[1] We model each individual as a self-interested reinforcement learning (RL) agent, adjusting its behavior so as to maximize the (individual) reward received throughout time. To incorporate the social signaling mechanism, we adopt the framework of intrinsically-motivated RL [3], [4] recently extended to multiagent scenarios [5], [6]. In our experiments, each agent is allowed to interact with another member of the population with a certain probability. For each agent, a *fitness-based* external reward related with resource consumption is combined with a *social-based* reward relating the internal status of the other interacting agent. To examine the impact of such social signals, we measure the combined performance of populations according to three distinct selection paradigms: one in which the fitness of the population is calculated by averaging the individual performances of all agents; another where only the most fit individual is considered; another where we consider the negative value of the variance between individual performances in the population.

The results from our experimental study show that this simple social signaling mechanism allows for the appearance of coordination strategies where the agents learn to "cherish" the welfare of others, especially when only half of the population may have access to the resources at each time. Another main result of the study is that this social mechanism induces social inequalities among the members of the populations, mainly when the resources are scarce. Such effect stems from the fact that healthier individuals gain a higher "status" when compared to other members, benefiting from their social displays being appreciated by others via the modeled social signaling mechanism.

II. Background

In this section we present the necessary technical background on the multiagent learning framework adopted in our experimental study.

A. Partially Observable Markov Games

To model a population of agents in limited resources scenarios we adopt the *partially observable Markov game* (POMG) model with K players.

A POMG proceeds as follows. At each time-step t, the game is in a state $S(t)$ from a finite set $\mathcal{S}$ of possible states. Each agent k has access to a partial view of $S(t)$, henceforth referred as the *observation*, denoted as $Z_k(t)$, of agent k at time-step t

This work was supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013.

[1] In this paper we model a simple and implicit signal-exchange mechanism that allows for cooperation in multiagent contexts. We refer to www.langev.com for a complete bibliographic archive concerning the evolution of explicit, language-based mechanisms using agents and robotic entities.

taking values in some finite observation space, $\mathcal{Z}_k$. We write $\mathsf{O}_k(z_k \mid s)$ to denote the observation probability $\mathsf{O}_k(z_k \mid s) \triangleq \mathbb{P}\left[Z_k(t) = z_k \mid S(t) = s\right]$.

At each time-step t, every agent k in the game selects, simultaneously and *without explicit communication*, an action $A_k(t)$ from a set $\mathcal{A}_k$ of possible actions. We write $A(t) = \langle A_1(t), \ldots, A_K(t)\rangle$ to denote the *joint action* of all agents at time-step t. The game then transitions, at time-step $t+1$, to a state $S(t+1) \in \mathcal{S}$ with a probability $\mathsf{P}(S(t+1) \mid A(t), S(t))$ that depends only on $S(t)$ and $A(t)$. Each agent k is awarded a reward $R_k(t+1)$ that depends only on $S(t)$ and $A(t)$ for some bounded real-valued (reward) function $r_k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Each agent k receives a new observation $Z_k(t+1)$ and the process repeats. The goal of each agent k is to select its actions so as to maximize the quantity

$$V_k(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_k(t+1) \mid S(0) = s\right], \tag{1}$$

where γ is a discount factor such that $0 \leq \gamma \leq 1$. We can compactly represent a K-player POMG as a tuple

$$\Gamma = (K, \mathcal{S}, (\mathcal{A}_k), (\mathcal{Z}_k), \mathsf{P}, (\mathsf{O}_k), (r_k), \gamma).$$

Solving a POMG amounts to compute, for each agent $k, k = 1, \ldots, K$, a mapping π_k, known as a *policy*, from each possible history of actions/observations,

$$H(t) = \{Z_k(0), A_k(0), \ldots, Z_k(t)\},$$

to a distribution over the set $\mathcal{A}_k$ of possible individual actions, in such a way that (1) is locally maximized for all agents (a situation known as *equilibrium*). Such policy is intractable to compute except for the simplest games [7].

Several specialized models, however, are amenable to efficient solutions. One such model is known as *Markov decision process* (MDP) [8], used to model single-agent settings where the agent has perfect perception of the state $S(t)$. Therefore, an MDP is a POMG where $K = 1$, $\mathcal{Z} = \mathcal{S}$, and O is the identity mapping.[2] An MDP can be compactly represented as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathsf{P}, r, \gamma)$ and, since an agent modeled as an MDP has (admittedly) perfect perception of the state $S(t)$ of the process, it is possible to find an optimal *deterministic* and *memoryless* policy, $\pi^* : \mathcal{S} \to \mathcal{A}$, such that

$$V^{\pi^*}(s) \triangleq \mathbb{E}\left[\sum_{t=0}^{\infty} R_k(t+1) \mid S(0) = s, A(t) = \pi(S(t))\right] \geq V^{\pi}(s),$$

for any policy π and any state $s \in \mathcal{S}$. In other words, in an MDP, it is possible to find a deterministic policy π^* that attains the maximum value for any initial state $S(0)$ and that depends only on the current state of the game. The function $V^* : \mathcal{S} \to \mathbb{R}$ defined as $V^*(s) = \max_\pi V^\pi(s)$ is such that $V^*(s) = V^{\pi^*}(s)$ and is known as the *optimal value function*. Similarly, π^* is known as an *optimal policy* and can easily be computed using dynamic programming, as long as P and r are known [8]. On the other hand, if P or r (or both) are unknown (in what is known as a *reinforcement learning*—or RL—problem), the agent may gather information by interacting with the environment and build (implicitly or explicitly) estimates $\hat{\mathsf{P}}$ and $\hat{r}$ of P and r that can then be used to compute π^*.

In this paper, we adopt a simple stochastic policy gradient algorithm that directly adjusts the probability of taking an action A_k given observation Z_k at time step t. Denoting

$$p_k(a \mid z) \triangleq \mathbb{P}\left[A_k(t) = a \mid Z_k(t) = z\right],$$

every time the agent k receives a reward r for executing $a \in \mathcal{A}_k$ after observing $z \in \mathcal{Z}_k$, we update the corresponding policy component as

$$p_k(a \mid z) \leftarrow p_k(a \mid z) + \alpha r, \tag{2}$$

where α is the learning rate. We then normalize the probabilities $p_k(\cdot \mid z)$ so that

$$\sum_{a \in \mathcal{A}_k} p_k(a \mid z) = 1.$$

B. Intrinsically Motivated Reinforcement Learning

In spite of their expressive power, the computational complexity involved in solving POMGs renders this model impractical for our purposes. In contrast, MDPs can be solved efficiently [9], although they are unsuitable to describe most multiagent problems of interest.

In this paper we are mainly interested in studying the emergence of cooperative behavior in groups of self-interested agents co-existing in a common environment.[3] As such, we consider *agents that reason about the world as if it were an MDP*, ignoring both the existence of other agents (inasmuch as that existence is not evident from their observations) as well as the possibility that their perception of the state $S(t)$ of the world may be imperfect. In general, the above approach will lead to a very crude model and have a significant impact in terms of the performance of the agent [11].

However, recent work on *intrinsically motivated reinforcement learning* (IMRL) [3], [4] showed that it is possible to use richer reward functions that implicitly encode additional information about the task of the agents to overcome some of its perceptual limitations, both in single-agent [3], [4], [12] and multiagent [5], [6] scenarios. In this framework, the performance of an RL agent given some task in a set of environments of interest provides a measure of the *fitness* of the agent. In this manner, distinct reward functions used by some agent account for different degrees of fitness.[4]

Formally, given a K-agent POMG, let

$$h_k(t) = \{z_k(0), a_k(0), \rho_k(1), \ldots, \rho_k(t), z_k(t)\}$$

denote the *history of interaction* of agent k with an environment up to time step t. The history corresponds to all

[2] Since there is a single agent, we can drop the subscript $_k$.

[3] Such agents can be seen as computational counterparts of simple fitness-maximizing individuals, and it may debatable whether such agents would reason extensively about the behavior of other agents in the environment [10].

[4] We note that in our experiments, *intrinsic motivations* refer to the extended repertoire of social rewards the agents have access to in the context of IMRL.

the information perceived by the agent directly from the environment, including all the observations and actions taken. $\{\rho_k(\tau), \tau = 1, \ldots, t\}$ corresponds to an "external" evaluation signal that, at each time step t, depends only on the underlying state $S(t-1)$ of the environment and the action $A_k(t-1)$ performed by the agent.[5] We write $h(t)$ to denote a particular realization of $H(t) = \langle H_1(t), \ldots, H_K(t)\rangle$, the *joint history of all agents* interacting in an environment.

Each agent is considered to be an *independent learner* [13], meaning that each learns its own policy based on its individual actions. Denoting by r_k the individual reward function used by agent k throughout time, we write $\mathbf{r} = [r_1, \ldots, r_K]$ to denote the vector of all reward functions and henceforth refer to $\mathbf{r}$ as the *POMG reward function*.

Given a particular finite joint history h, we measure the *combined performance* of a population of K agents as a function $f(h)$ of each agents' individual fitness $f_k(h_k), k = 1, \ldots, K$. In this paper, we consider the fitness of each agent as the summation in h_k of the external evaluation signal ρ_k, *i.e.*,

$$f_k(h_k) = \sum_{\tau} \rho_k(\tau). \tag{3}$$

We investigate the impact of social signaling mechanisms (by means of the reward) in terms of the performance of a population of agents as a whole.

III. Experimental Study

We now detail our experimental study to analyze the influence of social display signaling in competitive scenarios.

A. Scenario Description

In our experiments, we model a population of agents competing for limited resources in a shared environment using the POMG model described earlier. Each agent has two actions available, *i.e.*, $\mathcal{A}_k = \{Eat, Nothing\}$ that, if chosen by the agent at some time step t, respectively signals its intention towards eating or not a resource from the environment. Moreover, each agent k has a probability of meeting another agent ℓ denoted by $P_{meet}(k, \ell)$, $\ell \neq k$ selected randomly from the population. At each time step t, the state of each agent k is described by its hunger status, denoted by $U_k(t)$, which can have one of the three following values:

$$\{\text{FULL}_k, \text{ALMOST\_HUNGRY}_k, \text{HUNGRY}_k\}.$$

The state of the POMG is described by the combination of the hunger status of all agents, *i.e.*,

$$S(t) = \langle U_0(t), \ldots, U_K(t)\rangle.$$

Besides observing its own hunger status, each agent is able to observe, if a meeting with another agent ℓ occurs at time step t, whether that agent ℓ is hungry or not. This observation is denoted by $O_{k,\ell}(t)$ and can have one of the following values:

$$\{\text{NOTHING}, \text{OTHER\_HUNGRY}_\ell, \text{OTHER\_FULL}_\ell\},^6$$

where NOTHING denotes the event of not meeting another agent at some time step. Each agent's observation at time step t is then given by

$$Z_k(t) = (U_k(t), O_{k,\ell}(t)).$$

Whenever an agent eats, by choosing action Eat, its hunger status is reset to FULL. However, the environment in which the agents interact has only resources available for a predefined percentage of the population, denoted by P_{eat_max}. This means that, at each time step, only $P_{eat\_max} \times K$ agents or less are allowed to eat in order for them to become FULL. If an agent chooses action $Nothing$, or if too many agents chose action Eat at the same time, their hunger status subsequently increases, from FULL to ALMOST_HUNGRY and then to HUNGRY, where it remains until a resource is eaten.

Whenever an agent eats, its fitness is increased by a value of $\rho = 1/n_{eat}(t)$, where n_{eat}(t) represents the number of agents that chose action Eat at time step t. If an agent becomes HUNGRY, its fitness is decreased by an amount of $\rho = -1/K$. In all other cases, agents' fitness remains the same, *i.e.*, $\rho = 0$.

B. Social Signaling Mechanism

A simple approach to construct the set $\mathcal{R}$ of possible individual reward functions to be used in the multiagent IMRL approach is to take the linear span of some set Φ of real-valued reward features, $\Phi = \{\phi_1, \ldots, \phi_p\}$ [3], [12]. In the context of our study, we consider two reward features, $\Phi = \{\phi_{\mathtt{soc},k}, \phi_{\mathtt{fit},k}\}$, where

- $\phi_{\mathtt{soc},k}(s, a)$ can be interpreted as a measure of the social importance for agent k when the population executes the joint action a in state s. In our study, this feature is dependent on the interaction of agent k with another agent ℓ at time step t. It has a value of -1 whenever agent ℓ's state at time step $t+1$ is $U_\ell(t+1) = \text{HUNGRY}_\ell$, 1 when $U_\ell(t+1) \in \{\text{ALMOST\_HUNGRY}_\ell, \text{FULL}_\ell\}$, and 0 otherwise, *i.e.*, when no meeting occurs. In a sense, $\phi_{\mathtt{soc}}$ functions as a social display mechanism when two agents meet, rewarding actions possibly leading to the observation of "healthy companions" and penalizing decisions possibly leading others to starvation;
- $\phi_{\mathtt{fit},k}(s, a) = \mathbb{E}\left[\rho_k(t+1) \mid S(t) = s, A(t) = a\right]$ is agent k's fitness-based reward when the population performs the joint action a in state s.

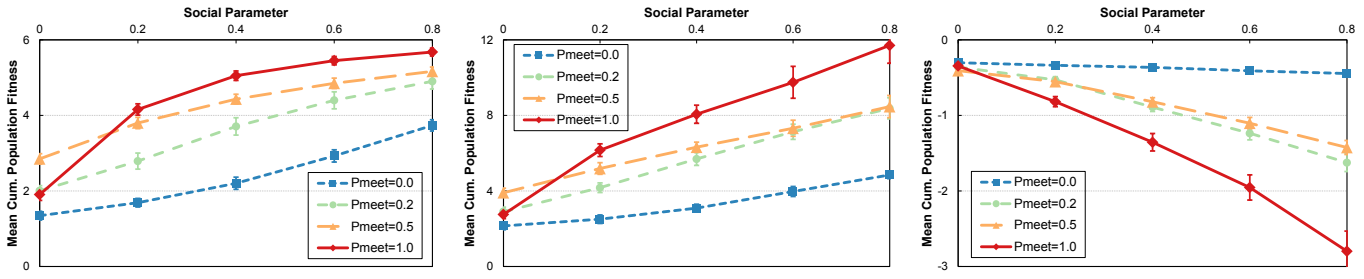
Each individual reward function can be defined as a linear combination of the features, *i.e.*, all functions in the form

$$r_k(s, a) = \boldsymbol{\phi}_k^\top(s, a)\boldsymbol{\theta}_k,$$

where $\boldsymbol{\phi}_k = [\phi_{\mathtt{soc},k}, \phi_{\mathtt{fit},k}]$ is agent's k feature vector and $\boldsymbol{\theta}_k = [\theta_{\mathtt{soc},k}, \theta_{\mathtt{fit},k}]$ is the vector of weights associated with

[5] In our experiments we consider this signal to be a kind of physiological feedback related with the agents' feeding behavior.

[6] For the purposes of our study, the ALMOST_HUNGRY hunger status of one agent is observed as being OTHER_FULL by another—this hunger status differentiation is important only for individual decisions by each agent.



(a) Avg. Pop. Fitness (b) Max. Pop. Fitness (c) Std. Dev. Pop. Fitness

Fig. 1. Combined performance of a population (the values) according to several meeting probabilities P_{meet} (the series) by varying the social parameter $\theta_{\texttt{soc}}$. We show the performance of each population according to the 3 fitness functions: (a) f_{avg}; (b) f_{max}; (c) f_{std_dev}. In all populations we set $P_{eat\_max} = 0.5$. Results correspond to averages over 60 independent Monte-Carlo trials.

each reward feature and corresponds to the parameters of the linear combination [3]. We write $r_k(\boldsymbol{\theta}_k)$ to explicitly denote the individual reward function corresponding to $\boldsymbol{\theta}_k$. In this paper, we study the emergence of cooperation in a population of siblings in the sense of [14], meaning that all agents learn with the same reward function parameterization, *i.e.*, $\boldsymbol{\theta}_k = \boldsymbol{\theta}, k = 1, \ldots, K$. The POMG reward function can thus be written as $\mathbf{r}(\boldsymbol{\theta})$, and $\boldsymbol{\theta}$ can be used as parameters of the population.

C. Population Fitness

In order to measure the combined performance of a population we adopt three different social paradigms that can be translated into different population fitness functions, namely $f_{avg}(h)$, $f_{max}(h)$ and $f_{std\_dev}(h)$, where

- $f_{avg}(h) = \frac{1}{K}\sum_{k=1}^{K} f_k(h_k)$ measures the fitness of a population during history h as the mean value of the individual fitness of each agent. f_{avg} acts as a selection function that chooses populations in which the average individual fitness is high;
- $f_{max}(h) = \max_k f_k(h_k)$ measures the fitness of the agent with the highest individual fitness during h, thus maximizing individual performance;
- $f_{std\_dev}(h) = -\sqrt{\frac{1}{K}\sum_{k=1}^{K}(f_k(h_k) - f_{avg}(h))^2}$ measures the fitness of a population as the negative standard deviation value of the individual fitness of all agents during history h. f_{std_dev} thus selects populations in which the individual fitness among all agents differs less.

D. Experimental Procedure

As we have seen from the scenario description above, both the "fitness-based" reward $\phi_{\texttt{fit},k}$ and the "social" reward $\phi_{\texttt{soc},k}$ depend on the joint actions taken by all agents. As such, the purpose of our experiments is to study the possible complex dynamics arising from the agents actions and see under which conditions the simple social signaling mechanism aids in the emergence of cooperation, and also how it varies according to the different fitness functions defined.

In our study, populations of interacting agents are described according to 4 different parameters, $\theta_{\texttt{soc}}$, $\theta_{\texttt{fit}}$, P_{meet} and P_{eat_max}, held fixed for all agents during each learning simulation. We sample the values for each parameter as follows:

- $\theta_{\texttt{soc}}$ is the social parameter associated with $\phi_{\texttt{soc}}$ and corresponds to the importance that each agent gives to the social displays of others. In our study, we sample $\theta_{\texttt{soc}}$ from the following values: $0, 0.2, 0.4, 0.6, 0.8$;
- $\theta_{\texttt{fit}}$ is the weight associated with $\phi_{\texttt{fit}}$ and corresponds to the importance that each agent gives to its own fitness. In each simulation, we set $\theta_{\texttt{fit}} = 1 - \theta_{\texttt{soc}}$;
- P_{meet} is the probability with which an agent observes another agent at each time step. In our experiments, we vary P_{meet} according to the following values: $0, 0.1, 0.2, \ldots, 1$;
- P_{eat_max} is the maximum percentage of the agents that are allowed to eat at each time step. In each simulation P_{eat_max} takes one of following values: $0.1, 0.2, \ldots, 1$.

We sampled a total of 3300 different populations corresponding to all the combinations of the parameters as described above. For each population, we simulated $K = 100$ agents competing for food resources in the environment and learning an individual policy as previously described. Finally, for each population we evaluate its combined performance according to the fitness functions detailed earlier.

IV. Results and Discussion

We now analyze the main results of our experimental study.

A. Social Influence on Population Fitness

The first thing we want to analyze is the impact of the social signaling mechanism on the combined performance of the population. Fig. 1 shows the change in performance of the population when we vary the social parameter ($\theta_{\texttt{soc}}$). We sample the performance for different values of meeting probability (P_{meet}) and fix $P_{eat\_max} = 0.5$, allowing half of the population to eat at each time step. The depicted results show that indeed the social signal mechanism has a positive impact on the population fitness, both when measuring the average of the individual performances (f_{avg}) and the maximal performance (f_{max}). This result thus shows that the signaling mechanism allows the agents to coordinate their feeding behavior. An examination to the individual policy of an average agent in these populations shows that it prefers to eat when HUNGRY, do nothing when FULL and have a mixed strategy in other cases. Overall, this kind of behavior allows

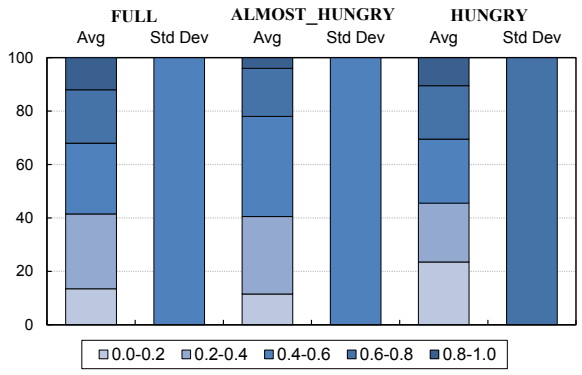


Fig. 2. Comparison between the distribution among agents of the probability of eating given the agent's hunger status, $p_k(Eat \mid U_{k(t)})$. Portions represent averages in relation to the observation of the other agent's status, $O_{k,\ell}(t)$. Darker portions mean higher probability of eating. The results correspond to the most fit population according to the function used. For f_{avg} (and also f_{max}), $\theta_{\texttt{soc}} = 0.8$ and $P_{meet} = 1$. For f_{std_dev} $\theta_{\texttt{soc}} = P_{meet} = 0$. In all populations, $P_{eat\_max} = 0.5$.

the agents to cooperate by feeding in turns and receiving positive social reward when letting other eat. Moreover, it is not surprising that the importance of the social signal increases with the probability of the agents meeting, *i.e.*, the more the agents meet, the faster their policy is adjusted towards the desired socially-aware behavior.

B. Social Inequalities

On the other hand, the results indicate that the higher the social parameter and the meeting probability, the more uneven the populations are in terms of individual fitness distribution, as depicted in Fig. 1(c), where f_{std_dev} decreases with $\theta_{\texttt{soc}}$ and P_{meet}. In fact, the population with the highest fitness shown in Fig. 1(a) ($\theta_{\texttt{soc}} = 0.8$ and P_{meet}=1) is the same as that shown in Fig. 1(b). Overall, this result shows that populations with socially-aware agents have an overall high combined performance and include also the most fit agents, but also allow for the greater inequalities in terms of individual performance among agents. Nevertheless, more "uniform" populations are comprised of "selfish" agents that care only for their individual fitness or agents that seldom meet, as depicted in 1(c) when $\theta_{\texttt{soc}} = 0$ or $P_{meet} = 0$, respectively.

To aid in this discussion, we depict in Fig. 2 the distribution among agents in a population of the eating probability given the hunger status. The results are taken from the most fit populations according to the several fitness functions. As we can see, the most fit socially-aware populations, as selected by f_{avg} (and f_{max}), are fairly composed of individuals using mix eating strategies. This means that although the agents all learn given the same parameters, the repeated interactions make the eating strategies to converge to distinct values, allowing for coordination but causing a large variance in the individual performances. On the other hand, "selfish" populations, as selected by f_{std_dev}, converge to a standardization in the agents' strategies towards eating—between $0.4 - 0.6$ when not HUNGRY and $0.6 - 0.8$ when HUNGRY. Ultimately, the similarities in the individual performances among the agents

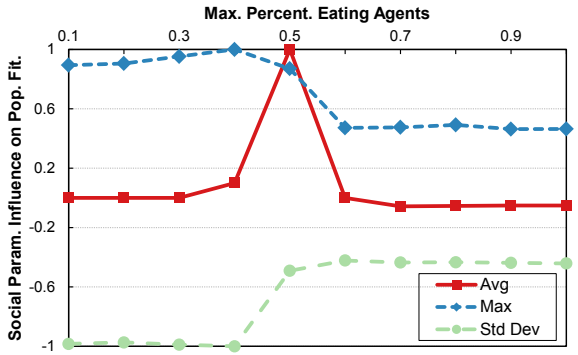


Fig. 3. Impact of the social parameter in the population fitness (the values) according to the 3 fitness functions (the series) by varying the maximum percentage of eating agents P_{eat_max}. In all populations we set $P_{meet} = 0.5$. Results correspond to averages over 60 independent Monte-Carlo trials.

mean that such populations are doomed to poor overall performances when compared with more socially-aware groups by not being able to learn to coordinate their eating behavior.

C. The Impact of Available Resources

In the previous subsections we analyzed the importance of the social signaling mechanism in the combined performance and fitness distribution within the population. We now analyze the influence of the social parameter according to changes in the amount of resources available. For that purpose, for each sampled value of P_{eat_max}, we normalize the fitness according to the difference between the performances of the population with the highest social influence ($\theta_{\texttt{soc}} = 0.8$) and the one with no social influence at all ($\theta_{\texttt{soc}} = 0$). As such, a positive value means that it compensates to be socially-aware, a value near 0 means that the social mechanism has very little impact on the population's performance, and a negative value translates to the social parameter having a negative impact, *i.e.*, the population is better off with no social signaling mechanism.

The results of this experiment are depicted in Fig. 3. As we can see, when measuring the average individual performance (f_{avg}), it only matters to pay attention to the social display of others when food resources are available for approximately half of the population ($P_{eat\_max} = 0.5$). This means that "sharing" does not matter when there are few resources available ($P_{eat\_max} < 0.5$), *i.e.*, when coordination is hard to achieve. On the other hand, if there are sufficient resources for the majority of the population ($P_{eat\_max} > 0.5$), then it also does not compensate too much to be socially-aware because even the "greedy" agents ($\theta_{\texttt{soc}} = 0$) are able to eat. The case is different when we use other metrics for the population's fitness. If we take the individual with the highest fitness (f_{max}), then it always compensates to be socially-aware, especially when there are fewer resources available ($P_{eat\_max} < 0.5$). This results shows that although it does not compensate to be socially-aware on average, using social signals makes for the emergence of some individuals with substantially higher fitness than the average population. In a sense, this interesting result shows the appearance of "popular" or "high-status" individuals that learn to benefit from being selfish and attain a

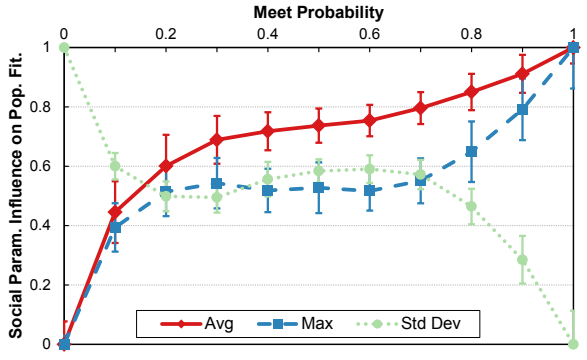


Fig. 4. Impact of the social parameter in the population fitness (the values) according to the 3 fitness functions (the series) by varying the meeting probability P_{meet}. In all populations we set $\theta_{\texttt{soc}} = 0.8$ and $P_{eat\_max} = 0.5$. Results correspond to averages over 60 independent Monte-Carlo trials.

higher fitness while others learn to appreciate "social reward" as provided by the healthy displays of the formers. This effect can also be observed from the mixed eating strategies found in the socially-aware population in Fig 2. As expected, the plot for the standard deviation case (f_{std_dev}) in Fig. 3 shows the symmetrical case, meaning that the populations with the most fit individuals are the ones with higher inequalities among agents, the worst case being when there are no resources for the majority of the population ($P_{eat\_max} < 0.5$).

D. The Impact of Social Encounters

Let us now study how the number of social encounters impacts the combined performance of the most socially-aware population, *i.e.*, when $\theta_{\texttt{soc}} = 0.8$. For that purpose, we plot in Fig 4 the fitness attained by a population according to the three possible functions, normalized across all values of P_{meet}. Given the impact of the available resources on the populations' fitness, we set $P_{eat\_max} = 0.5$. The results of this experiment show that, when measuring the average (f_{avg}) or maximal (f_{max}) individual performance, the more the social encounters the higher the population's fitness. This means that highly social individuals benefit from interacting with other agents, allowing for a faster learning of the coordination strategy. On the other hand, the results show that the social encounters may be used by selfish agents to gain fitness over other members of the population, increasing the proportion of the resources eaten by a few agents as the number of encounters increases.

V. Conclusions

In this paper we analyzed the impact of simple social signaling mechanisms in the combined performance a population of agents in the context of multiagent competitive learning. In a simulated environment, self-interested reinforcement learning agents interact and compete with each other for limited resources. We followed previous works within multiagent IMRL and provided the agents with a social signaling mechanism that influences the amount of reward received throughout time. In a series of experiments, we studied the change in the fitness of a population of agents induced by varying the amount of resources available in the environment, the probability of two agents interacting, and the importance given to the social display of others. The combined performance of the populations was measured according to different metrics emulating distinct social paradigms.

Overall, the results from our experimental study suggest that given the nature of the competitive environment modeled, the social signaling mechanism allows both for the appearance of coordination strategies—allowing agents to learn socially-aware behaviors that trade short-term benefits for social acknowledgment from other members—, but also more social inequalities—leading to the emergence of highly-praised individuals that benefit from the intrinsic value provided to other members by displaying their full-satiation status.

In the future we would like to test the emergence of the importance of social signaling in heterogeneous populations in which each agent learns according to its own (different) social parameterization. For this purpose, one could provide the agents a simple imitation mechanism that would change the social parameterization according to the differential of fitness between interacting individuals. Such mechanism would afford the appearance of distinct personalities within the population, and we could thus examine the evolution of such individual traits in a more natural manner. Another interesting analysis would consist in examining the creation of subgroups of agents within the population sharing a similar personality.

References

[1] B. Skyrms, *Signals: Evolution, Learning and Information*. Oxford University Press, 2010.
[2] D. K. Lewis, *Convention: A philosophical study*. Cambridge, MA: Harvard University Press, 1969.
[3] S. Singh, R. Lewis, A. Barto, and J. Sorg, "Intrinsically motivated reinforcement learning: An evolutionary perspective," *IEEE Trans. Autonomous Mental Development*, vol. 2, no. 2, pp. 70–82, 2010.
[4] S. Singh, R. Lewis, and A. Barto, "Where do rewards come from?" in *Proc. 31st Annual Conf. Cognitive Science Society*, 2009, pp. 2601–2606.
[5] P. Sequeira, F. S. Melo, R. Prada, and A. Paiva, "Emerging Social Awareness: Exploring Intrinsic Motivation in Multiagent Learning," in *Proc. 1st Joint IEEE Int. Conf. on Development and Learning and on Epigenetic Robotics*, ser. ICDL-EPIROB 2011. IEEE, 2011.
[6] B. Liu, S. Singh, R. L. Lewis, and S. Qin, "Optimal rewards in multiagent teams," in *Proc. 2nd Joint IEEE Int. Conf. on Development and Learning and Epigenetic Robotics*. IEEE, Nov. 2012, pp. 1–8.
[7] J. Goldsmith and M. Mundhenk, "Competition adds complexity," in *Adv. Neural Information Proc. Systems*, vol. 20, 2007.
[8] M. Puterman, *Markov Decision Processes*. Wiley-Interscience, 1994.
[9] M. Littman, T. Dean, and L. Kaelbling, "On the complexity of solving Markov decision problems," in *Proc. 11th Int. Conf. Uncertainty in Artificial Intelligence*, 1995, pp. 394–402.
[10] R. M. Axelrod, *The evolution of cooperation*. New York, USA: Basic Books, Inc., Publishers, 1984.
[11] M. Littman, "Memoryless policies: Theoretical limitations and practical results," in *From Animals to Animats 3: Proc. of the 3rd Int. Conf. on Simulation of Adaptive Behavior*, 1994, pp. 238–247.
[12] J. Sorg, S. Singh, and R. Lewis, "Internal rewards mitigate agent boundedness," in *Proc. 27th Int. Conf. Machine Learning*, 2010, pp. 1007–1014.
[13] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," in *Proc. 15th Nat. Conf. Artificial Intelligence*, 1998, pp. 746–752.
[14] T. Bergstrom, "On the evolution of altruistic ethical rules for siblings," *American Economic Review*, vol. 85, no. 1, pp. 58–81, 1995.