Emerging Social Awareness: Exploring Intrinsic Motivation in Multiagent Learning

Pedro Sequeira, Francisco S. Melo, Rui Prada and Ana Paiva
Instituto Superior Técnico / INESC-ID
Av. Prof. Dr. Cavaco Silva
2744-016 Porto Salvo, Portugal
Email: pedro.sequeira@gaips.inesc-id.pt, fmelo@inesc-id.pt, rui.prada@gaips.inesc-id.pt, ana.paiva@inesc-id.pt

***Abstract*—Recently, a novel framework has been proposed for intrinsically motivated reinforcement learning (IMRL) in which a learning agent is driven by rewards that include not only information about what the agent must accomplish in order to "survive", but also additional reward signals that drive the agent to engage in other activities, such as playing or exploring, because they are "inherently enjoyable". In this paper, we investigate the impact of intrinsic motivation mechanisms in multiagent learning scenarios, by considering how such motivational system may drive an agent to engage in behaviors that are "socially aware". We show that, using this approach, it is possible for agents to *learn individually* to acquire *socially aware behaviors* that trade-off individual well-fare for social acknowledgment, leading to a more successful performance of the population as a whole.**

I. Introduction

One fundamental skill expected of an intelligent agent is that it should be able to autonomously *learn* how to perform new tasks and behave in situations never experienced before. And, while the discipline of machine learning proposes a range of possible answers to the question of "how" to learn, it does not provide any satisfying answer to "why" an agent should learn. Endowing an artificial system such as a robot with a self-motivated, open-ended system for learning increasingly complex behaviors is therefore a fundamental challenge in artificial intelligence and robotics.

Intrinsic motivation is a term coined in the psychology literature, and refers to the "forces" that drive organisms to engage in certain activities because they are inherently enjoyable, such as playing, exploring, etc. [1]. Several parallels have been drawn between intrinsic motivation systems in the psychology literature and some active learning and experimental design techniques from machine learning [2]–[4]. Also, several intrinsic motivation systems have been proposed for artificial systems in areas such as developmental robotics. Examples include the hierarchical acquisition of skills using intrinsically motivated reinforcement learning [5]–[7], the acquisition of a visual-attention system from motivation variables [8], and others [3], [4], [9], [10].

Recently, a novel framework has been proposed for intrinsically motivated reinforcement learning (IMRL) [11], [12]. This framework proposes an *evolutionary interpretation* of intrinsic rewards, according to which the latter are the result of an (evolutionary) optimization process that maximizes the *expected fitness* of an agent given some distribution of environments of interest. Within this framework, an agent's reward signals include information about what the agent must accomplish in order to "survive"—the so-called *extrinsic* reward signals—but also additional reward signals that drive the agent to engage in other activities such as playing or exploring—the *intrinsic* reward signals. How each reward signal contributes to the agent's overall reward is "hard-wired" and depends on the range of environments that the agent is expected to interact with and on how "fitness" is measured. The optimization of such contributions can be interpreted as the outcome of an evolutionary process that conditions a particular agent interacting with and learning in certain environments to weight different reward signals in a way that is best for this agent and these environments.

From a more computational perspective, it was shown that optimized rewards obtained using this approach include information that allows agents to overcome limitations such as perceptual aliasing [13]. Also, as argued in [12], [13], the mechanism of reward optimization is fundamentally different from *reward-shaping* [14], as the latter approach does not modify the optimal policy for a certain environment. In environments where the limitations of an agent lead to optimal policies which can perform poorly, reward shaping cannot be expected to allow the agent to overcome such limitations. The use of intrinsic reward signals, on the other hand, can lead to a tremendous boost in performance, since it actually does modify the optimal policy. The IMRL framework provides a computationally sound approach to the implementation of intrinsic motivation systems in learning agents.

In this paper, we explore the impact of such motivational systems in multiagent scenarios. While most aforementioned works focus on single-agent scenarios, in which the motivational system drives a single agent to engage in behaviors that are not directly "survival"-related, in this paper we consider how these same systems may drive an agent to engage in behaviors that are "socially aware", in a sense to be made clear. Specifically, we adopt the IMRL framework of [11], [12] and show that, using this approach, it is possible for agents that *learn individually* to acquire *socially aware behaviors*

This work was partially supported by the Portuguese Fundação para a Ciência e a Tecnologia (INESC-ID multiannual funding) through the PIDDAC Program funds. The first author acknowledges the PhD grant SFRH/BD/38681/2007 from the Fundação para a Ciência e a Tecnologia.

978-1-61284-990-4/11/$26.00 ©2011 IEEE

that trade-off individual well-fare for social acknowledgment, leading to more successful populations. In a sense, we show that the IMRL framework can be used to endow agents with *social motivation* that drives them to learn such behaviors.

Given the evolutionary interpretation proposed for the IMRL framework that we adopt [12] and the social nature of the scenarios we consider, there is a close relation between our work and *evolutionary game theory* [15]. In its simplest form, evolutionary game theory analyzes the dynamics of a population when invaded by a small group of mutants. The interaction between individuals (mutant or non-mutant) is modeled as a strategic game in which the payoffs measure the fitness of the individuals after the interaction. Game theoretic notions are then used to analyze the dynamics of the fitness of the population and predict whether the mutants will eventually extinguish or not [15]. In our setting, the overall reward obtained after the reward optimization process can be seen as a *evolutionary stable equilibrium* in a sibling population, as discussed in [16].[1]

The paper is organized as follows. Section II discusses the general reinforcement learning setting and the IMRL approach from [12]. Section III presents the application of IMRL framework in multiagent scenarios and discusses possible reward signals that can be considered as intrinsic rewards. Section IV illustrates the application of this framework to multiagent systems and discusses some of the socially-aware behaviors observed. Section V provides some conclusions discussing some possible improvements and future work

II. Background

In this section we introduce the fundamental RL concepts and review the IMRL framework used in this paper [12].

A. Reinforcement Learning

Reinforcement learning (RL) addresses the general problem of an agent faced with a sequential decision problem [17]. By a process of trial-and-error, the agent must learn a "good" mapping that assigns perceptions to actions. Such mapping determines how the agent acts in each possible situation and is commonly known as a *policy*.

We model the sequential decision problem faced by an agent as a *partially observable Markov decision process* (POMDPs), denoted by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{Z}, \mathsf{P}, \mathsf{O}, r)$. At every step t, the state can be in any of a finite set $\mathcal{S}$ of possible states. Depending on its current perception z_t of the state, the agent chooses an action a_t from a finite set of possible actions, $\mathcal{A}$, and the environment transitions from state s_t to state s_{t+1} with probability $\mathsf{P}(s_{t+1} \mid s_t, a_t)$. The agent receives a reward $r(s_t, a_t)$ and makes a new observation z_{t+1} from a set of possible observations, $\mathcal{Z}$, with probability $\mathsf{O}(z_{t+1} \mid s_{t+1}, a_t)$, and the process repeats. The goal of the agent is to choose its actions so as to gather as much reward as possible, discounted by a positive discount factor $\gamma < 1$. Formally, this corresponds to maximizing the value

$$v = \mathbb{E}\left[\sum_t \gamma^t r(s_t, a_t)\right]. \tag{1}$$

The reward function r implicitly encodes a *task*, which the agent must complete by finding a *policy* $\pi^* : \mathcal{Z} \to \mathcal{A}$ that maximizes the value in (1). In typical RL scenarios, it is assumed that observations z_t correspond to the actual states s_t of the agent/environment [17]. When this is the case, it is possible to find a *policy* $\pi^* : \mathcal{Z} \to \mathcal{A}$ maximizing the value in (1). Such optimal policy can be derived from the *optimal Q-function*, Q^*, that determines how good (in the long-run) each action $a \in \mathcal{A}$ is, in each state $s \in \mathcal{S}$, if the agent performs optimally afterwards. In other words,

$$\pi^*(s) = \operatorname*{argmax}_{a \in \mathcal{A}} Q^*(s, a).$$

This function can be learned using any of a number of methods [17]. In this paper, our learning agents run the Dyna-Q/prioritized sweeping algorithm [18], where the observations of the agent are treated as states. In Dyna-Q, the agent uses its interaction with the environment to construct a model $\hat{\mathcal{M}}$ of the MDP $\mathcal{M}$ and uses this model to compute Q^*.

B. Intrinsically Motivated RL

In the IMRL framework, learning agents are evaluated according to their *expected fitness* throughout their lifetime. The learning agent is expected to interact with one among a set $\mathcal{E}$ of possible environments, and optimize its policy with respect to one among a set $\mathcal{R}$ of possible rewards. Depending on the specific environment and reward, the agent produces a history h and its fitness $\mathcal{F}(h)$ is evaluated with respect to this history by some given fitness function $\mathcal{F}$. An optimal reward function $r^* \in \mathcal{R}$ is such that the expected fitness of the agent with respect to a distribution over possible environments is maximized.

In this paper, we consider $\mathcal{R}$ as the set of all rewards of the form

$$r(s, a) = \sum_i \theta_i \phi_i(s, a, h), \tag{2}$$

where each ϕ_i is referred as a *reward signal*. The weights θ_i determine the contribution of these reward signals to the overall reward that the agent will learn to maximize throughout its lifetime. We also admit the reward signals ϕ_i to be *history-dependent* (and thus, non-Markovian). As in [12], this dependence is not considered by the learning algorithm, but does impact the corresponding optimal policy.

We refer to a *fitness-based reward signal* as a reward-signal $\phi^{\mathcal{F}}$ that explicitly rewards fitness-maximizing states [12]. For ease of exposition, we henceforth refer to such reward signal as the *extrinsic reward*, and the other reward signals as *intrinsic rewards*, although these designations may not correspond to any such quantities found in biological systems. We denote the weight-vector corresponding to the optimal reward function as $\boldsymbol{\theta}^*$. This vector can be computed in any of a number of ways

[1] Due to space limitations, a detailed discussion of the relation between evolutionary game theory and our approach is out of the scope of this paper.

[12], [19]. The particular method considered is not important for our purposes, so we adopt a simple approach in which the weights are optimized by a brute-force search in the weight-space.

III. Socially-Motivated Learning Agents

In this section we apply the IMRL approach to multiagent scenarios. In particular, we discuss two (social) reward signals to be used within the IMRL framework and how these relate to specific social interactions studied in the specialized literature.

A. Fitness in Multiagent Scenarios

We consider a scenario where N agents interact in a common environment, among a set of $\mathcal{E}$ possible environments. The evolution of the state of the environment generally depends on the actions of *all* agents, and each agent k, $k = 1, \ldots, N$, has access to a local observation function O_k that maps the state of the environment into a local observation z_k. Each agent must learn to optimize its individual policy π_k with respect to an individual reward function r_k, chosen among a set $\mathcal{R}$ of possible reward functions that take the general form described in (2). Each agent k, $k = 1, \ldots, N$, is evaluated according to its expected fitness $\mathcal{F}_k$, and the overall population is evaluated according to their *summed fitness*,

$$\mathcal{F}(\mathbf{h}) = \sum_{k=1}^{N} \mathcal{F}_k(h_k),$$

where we denote by h_k the history of agent k and by $\mathbf{h}$ the joint history of all N agents, *i.e.,* $\mathbf{h} = (h_1, \ldots, h_N)$. For simplicity, we focus mostly on homogeneous scenarios, where all agents share similar reward functions and observation functions.

We note that our agents are *individual learners* in the sense of [20]. Therefore, they do not explicitly reason about other agents or their course of action. It follows that, whatever "collaborative" behavior emerges from the interaction among the agents, it does not result from any explicit social considerations that the agents may have on the well-fare of others.

B. Social Reward Signals and Social Motivation

To develop reward signals that take into consideration social interactions among agents, we consider the notion of *affiliation* from Dörner's PSI-theory [21], later adapted to the PSI agent architecture [22]. PSI agents have an urge to affiliate with other agents by sending and receiving *legitimacy signals*, also known as *l-signals*, that reward successful interactions. Dörner further defines other social signals that facilitate social interactions: *anti-l-signals* that punish unsuccessful social interactions, and *internal l-signals* that reward socially-acceptable behaviors [21], [22].

In our IMRL framework, we use reward signals that can be interpreted as computational counterparts to the several social signals discussed above, and show that these signals lead to the emergence of "socially aware" behaviors in social contexts. As will soon become apparent, the use of these signals as intrinsic rewards improves the overall fitness of the agent population without significantly impacting the individual fitness of each agent within that group.

We propose two possible social reward signals. One can be interpreted as a computational counterpart to the internal l-signals, and is hereby denoted by ϕ^i. The second can be interpreted as a computational counterpart to the (external) l-signals, and is hereby denoted as ϕ^e. These reward signals can be interpreted as representing how satisfied the affiliation need of an agent is. In our IMRL framework, the individual rewards r_k are thus given by

$$r_k(s, a) = \theta^{\mathcal{F}} \phi^{\mathcal{F}} + \theta^e \phi^e + \theta^i \phi^i,$$

where, as seen before, $\phi^{\mathcal{F}}$ denotes the fitness-based reward signal, or extrinsic reward. The weights $\theta^{\mathcal{F}}$, θ^e and θ^i are scalar values between 0 and 1 that indicate the contribution of each reward signal, $\phi^{\mathcal{F}}$, ϕ^e and ϕ^i, to the overall reward that the agents will learn to maximize throughout their lifetime. For example, a weight vector $\boldsymbol{\theta} \triangleq [\theta^{\mathcal{F}}, \theta^e, \theta^i] = [1, 0, 0]$ corresponds to an agent that values only the extrinsic reward signal, $\phi^{\mathcal{F}}$, while completely ignoring the social reward signals.

In accordance with the IMRL framework [12], the weight vector $\boldsymbol{\theta}$ are optimized for all agents to maximize the average fitness $\mathcal{F}$ of the population. The optimal weight vector $\boldsymbol{\theta}^*$ relates to the way that natural agents are phylogenetically predisposed to socially behave within a particular population. For example, some species may favor fair resource sharing while others may live within a highly hierarchical social structure that favors an unbalanced resource distribution within its members. Each weight configuration yields different degrees of fitness for the particular population and range of environments in which the agents co-exist.

C. Social Reward Signals in Limited Resource Scenarios

To apply the ideas discussed so far in a concrete multi-agent scenario, we resort to an adaptation of the foraging environments described in [12]. Our purpose is to illustrate the emergence of socially-aware behaviors and investigate the impact that such behaviors can have in the fitness of the overall population, comparing it with that of a population of "selfish" agents. For simplicity of presentation and analysis, we focus on 2-agent scenarios.

In our scenarios, the agents have limited food resources available in the environment, and the legitimacy signals are used to provide "social feedback" on their feeding behavior. Specifically, we assume that the agents know who was the last agent to consume a food resource from the environment, and are able to detect another agent when they are co-located in the environment. In all our experiments, we consider

$$\mathcal{F}_k(h_k) = \sum_{h_k} \phi_k^{\mathcal{F}},$$

i.e., the fitness of agent k corresponds to the total extrinsic reward received by agent k throughout its lifetime. We also define the following events, used in the calculation of the intrinsic reward signals:

- $\text{LTE}_k(t)$ denotes the event that, at step t, agent k was the last agent to consume a food resource;
- $\text{FOOD}_k(t)$ denotes the event that, at step t, agent k is located near a food resource;
- $\text{FULL}_k(t)$ denotes the event that, at step t, agent k is in a fully satiated state;
- $\text{HUNGRY}_k(t)$ denotes the event that, at time step t, agent k is in a hungry state; We note that each agent cannot be hungry and full at the same time, *i.e.,* $\forall_t \text{FULL}_k(t) \uparrow \text{HUNGRY}_k(t)$, but they can be in an intermediate *satisfied* state, when they are neither full or hungry;
- $\text{EAT}_k(t)$ denotes the event that, at step t, agent k took action "Eat" (*i.e.,* tried to consume a food resource);
- $\text{OTHER\_EAT}_k(t)$ denotes the event that, at step t, the agents are co-located and the other agent took action "Eat".

All above events can be perceived by agent k from its observation z_k at time t. Using the above events, we can now define the reward signals ϕ^e and ϕ^i and discuss their relation with internal and external l-signals.[2]

In our scenario, consuming food is the only behavior that directly contributes to the fitness of the population, while being hungry decreases the fitness. Therefore,

$$\phi^{\mathcal{F}}_k(t) = \mathbb{I}\left[\text{FULL}_k(t)\right] - 0.15 \cdot \mathbb{I}\left[\text{HUNGRY}_k(t)\right]$$

where $\mathbb{I}[e]$ denotes the indicator function for event e. Moreover, ϕ^e_k and ϕ^i_k reward and punish the feeding behavior of agent k depending on whether it is or not socially aware, *i.e.,* if it takes into consideration whether it was the last agent to eat. Formally, we define the external social reward signal as

$$\phi^e_k(t) = \mathbb{I}\left[\text{LTE}_k(t)\right] \cdot \mathbb{I}\left[\text{FOOD}_k(t)\right] \cdot \mathbb{I}\left[\text{OTHER\_EAT}_k(t)\right] \cdot \left(\mathbb{I}\left[\neg\text{EAT}_k(t)\right] - \mathbb{I}\left[\text{EAT}_k(t)\right]\right),$$

where $\neg\text{EAT}_k$ denotes the event that agent k did not take the action "Eat"; and the internal social reward signal as

$$\phi^i_k(t) = \mathbb{I}\left[\text{LTE}_k(t)\right] \cdot \mathbb{I}\left[\text{FOOD}_k(t)\right] \cdot \left(\mathbb{I}\left[\neg\text{EAT}_k(t)\right] - \mathbb{I}\left[\text{EAT}_k(t)\right]\right).$$

Informally, ϕ^e_k rewards agent k for allowing the other agent to eat when k was the last to eat, and punishes agent k for eating in the presence of the other agent when k was the last to eat. Conversely, ϕ^i_k rewards agent k for not eating when it was the last to eat, and punishes agent k when it eats having been the last to eat.

Let us now consider the relation between the above reward signals and the legitimacy-signals from the PSI-theory:

- *l-signals* and *anti-l-signals* somehow encode the degree of acceptance of the conduct of an individual by other members of its social group [22, p. 128]. In our setting, the reward signal ϕ^e_k takes a positive value whenever agent k has the possibility to consume a food resource but chooses not to because it was the last to eat, thus giving the other agent the opportunity to be satiated. Since this corresponds to a sensible behavior, the positive value of ϕ^e_k can be seen as a signal of acceptance by the other agent. A similar argument can be drawn with respect to the negative values of ϕ^e_k.
- *internal l-signals* measure how much an agent's actions are in accordance of its own *internal standards* [22, p. 128]. In our setting, the reward signal takes a positive value whenever agent k has the possibility to consume a food but decides not to because it was the last to eat (independently of the presence or absence of other agents). Because our scenarios deal with food sharing, this reward signal somehow encodes the degree of satisfaction that agent k gets for engaging in such *altruistic* behavior.
  Socially-aware agents feel *intrinsically rewarded* when they feel they engage in a behavior for the benefit of other members of its social group [23, p. 281]. Moreover, altruistic behaviors may carry an initial cost that is only compensated after a certain time period [23, p. 281]. In our setting, ϕ^i_k rewards agent k for altruistic behavior, even if this implies smaller individual fitness. Similarly, it punishes selfish behaviors, even if the agent was not directly competing with another agent for a food resource.

[2]Of course that we provide only one possible interpretation of these quantities in the context of our very simple scenario, and do not claim this interpretation to be universal or even biologically plausible. However, as will become apparent, this interpretation facilitates the discussion of the observed results. Depending on the social environments being considered, other features may be defined.

IV. Experiments and Results

A. Experimental Setup

In our experiments, two learning agents co-exist in the environment depicted in Fig. 1. Each agent has available 5 possible actions, $\{N, S, E, W, \text{Eat}\}$. The four direction actions move the agent deterministically to the adjacent cell in the corresponding direction; the action Eat consumes a food resource if one is present in the agent's location, and does nothing otherwise. At each time step, the agent observes its current position, its satiation status (*i.e.,* whether it is HUNGRY, SATISFIED or FULL), and whether food or another agent are present at the agent's current location. It also knows whether it was the last one to eat. Whenever an agent consumes a food resource, it becomes FULL for one time-step, after which it returns to the SATISFIED state. If it does not consume any resources for 30 time-steps, it becomes HUNGRY. As already seen, the extrinsic reward $\phi^{\mathcal{F}}$ of each agent k at time t depends only on the hunger-status of the agent, and is given by $\phi^{\mathcal{F}}_k(t) = \mathbb{I}\left[\text{FULL}_k(t)\right] - 0.15 \cdot \mathbb{I}\left[\text{HUNGRY}_k(t)\right]$.

We ran three different experiments, each consisting of a variation of the general problem defined above. These experiments differ in the amount of food resources available, the start position of the agents (see Fig. 1 for specific cell positions) and particular distinctions in the process of feeding.

- *Single-food scenario*: In this scenario there is always one food resource available at $(3 : 3)$. One of the agents departs from position $(1 : 3)$ while the other departs from $(8 : 3)$. Whenever one agent consumes the food

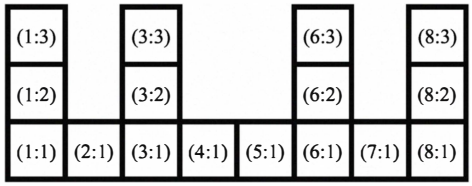


Figure 1. The foraging environment used in the experiments. Each square marked by $(x : y)$ coordinates represents a possible location for the agent.

resource it is repositioned in $(1 : 3)$, while the other agent is repositioned at $(8 : 3)$. The placement of the agents gives an advantage to the last eating agent, as it can reach the food source faster and allow the other to starve. Whenever the two agents try to consume the resource simultaneously, neither of them succeeds.

- *Equal-resource scenario*: In this scenario there are always food resources available at $(3 : 3)$ and $(6 : 3)$. As before, one of the agents departs from $(1 : 3)$ while the other departs from $(8 : 3)$. Also, whenever one agent consumes a resource, it is repositioned in $(1 : 3)$ and the other agent is repositioned at $(8 : 3)$. Whenever the two agents try to consume the same resource simultaneously, neither of them succeeds. However, the fact that two food resources are available allows both agents to eat simultaneously.
- *Stronger-agent scenario*: In this scenario there is always one food resource available at $(3 : 3)$. Both agents depart from position $(8 : 3)$. Whenever one agent consumes the food resource they are both repositioned in $(8 : 3)$. In this scenario, however, Agent 1 is stronger than Agent 2 and whenever the two agents try to consume the resource simultaneously, only Agent 1 succeeds. This gives an advantage to Agent 1, as it can always overpower the other and allow the other to starve.

From each agent's perspective, the environments are non-Markovian, since there are elements of the state that the agents cannot observe (*e.g.,* the position of the other agent). In our Dyna-Q implementation, we use a learning rate $\alpha = 0.3$ and $\gamma = 0.9$. The agents follow an ε-greedy policy with a decaying exploration rate $\varepsilon_t = \lambda^t$, with $\lambda = 1.00005$. To optimize the reward function, we adopt an adaptive sampling approach similar to the one in [12]. The optimization process determines, for each scenario, the optimal weight vector $\boldsymbol{\theta}^*$ that maximizes the overall fitness of the population.

B. Results

We simulated our agents for 100 000 learning steps and present in Table I the overall fitness obtained in each of the test scenarios. The results correspond to averages of 100 independent Monte-Carlo trials.

For each scenario, we compare the fitness of the population consisting of two agents with optimized weight vectors with that of a population consisting of two standard Dyna-Q agents that learn using only the extrinsic reward. As can be seen from the results in Table I, socially motivated agents attain a greater fitness than the population using only the extrinsic reward

Table I
POPULATION FITNESS FOR EACH SCENARIO. THE FIRST COLUMN INDICATES THE OPTIMAL WEIGHT VECTOR FOR EACH ENVIRONMENT. THE COLUMN MARKED "OPTIMAL" CORRESPONDS TO THE POPULATION WITH THE OPTIMIZED WEIGHT VECTOR; THE COLUMN MARKED "EXTRINSIC" CORRESPONDS TO THE STANDARD DYNA-Q AGENTS ($\boldsymbol{\theta} = [1, 0, 0]$).

Scenario	$\boldsymbol{\theta}^* = [\theta^{\mathcal{F}}, \theta^e, \theta^i]$	Optimal	Extrinsic
Single-food	$\boldsymbol{\theta}^* = [0.33, 0.33, 0.33]$	2 355.8	−4 029.9
Equal-resource	$\boldsymbol{\theta}^* = [0.20, 0.80, 0.00]$	13 969.9	13 629.1
Stronger-agent	$\boldsymbol{\theta}^* = [0.14, 0.57, 0.29]$	8 019.2	441.7

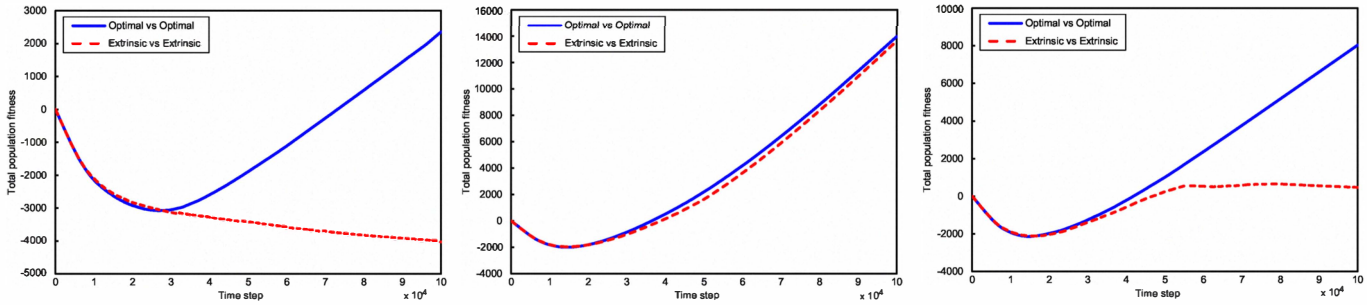
during learning. These agents are typically "selfish" which, combined with the structure of the environments, typically leads one of the agents to starvation.

In the Single-food scenario, the optimal weight vector fairly distributes the importance of the three reward signals considered. This indicates that the proposed reward signals, inspired by social legitimacy-signals, do provide a relevant trade-off between the fitness-based signal and the social reward signals. Each agent takes into consideration not only the extrinsic reward from being satiated, but are also sensible to the social "reinforcement" received for allowing food sharing. If this was not the case, then the last agent to eat could easily improve its individual fitness by starving the other agent. This can be seen from the results achieved using only the extrinsic reward, corresponding to the last column of Table I.

The Equal-resource scenario provides insights on situations where food resources are always abundant. In this case, there needs to be no food sharing or competition over the available resources. As a consequence, the optimal weight vector for this scenario completely ignores ϕ_k^i while giving more attention to the extrinsic reward provided by eating food. As expected, the results in terms of fitness achieved by either the optimal and the extrinsic weight sets are very similar. This is due to the fact that, by having ***unlimited*** food resources available to both agents, they don't have to signal each other for acceptable feeding behaviors. However, this does not mean that they should ignore all intrinsic motivation coming from the social features. For example, due to the placement policy after eating, in the beginning the agents might learn to obtain reward by eating the food resource at $(3 : 3)$ while ignoring the fact that there is also a resource available at $(6 : 3)$. In such cases, as the optimal weight set indicates, ϕ_k^e plays an important role by reprehending selfish behaviors during competition.

Finally, in the stronger-agent scenario, our objective was to see whether the social behavior arises out of a ***need*** to cooperate with each other to avoid starvation if one agent gets to be the last to learn. In this scenario, by letting both agents depart from the same position and having one agent always overpower the other, Agent 1 strictly has no need to cooperate with Agent 2. However, even in this situation, we observe that socially-aware behavior emerges, leading to resource sharing. This can also be seen from the optimal weight vector for this scenario, which places significant weight in both ϕ^e and ϕ^i.

We conclude this section by illustrating in Fig. 2 the evolution of the fitness of the population in each of the



(a) Single-food scenario (b) Equal-resource scenario (c) Stronger-agent scenario

Figure 2. Evolution of the fitness of the population in each of the three test scenarios. Results are averages over 100 independent Monte-Carlo trials. "Optimal vs Optimal" corresponds to populations of two agents with the optimal weight vectors. "Extrinsic vs Extrinsic" corresponds to populations of two agents with the weight vector $\boldsymbol{\theta} = [1, 0, 0]$.

different scenarios.

V. Conclusions

In this paper we investigated the impact that a social motivation system can have in the emergence of socially aware behaviors in a population of learning agents. We adapted the framework for intrinsically motivated reinforcement learning in [12] and explored how simple social signals inspired by the notion of affiliation proposed in the PSI-theory. Our results show that, indeed, our socially motivated agents perform as a whole much better than "selfish" agents, with little impact in their individual fitness. Our results show that, even in the presence of dominating agents, *i.e.,* agents who do not require socially aware behavior to maximize their individual fitness, socially aware behaviors lead to an improved population fitness. Finally, our results that the social motivation does not blindly lead to selfless behavior: in scenarios where resources abound, our agents learn to disregard the needs of the others, since they are not affected by the agent's behavior.

While the purpose of this work is to investigate the use of intrinsically motivated reinforcement learning in multiagent systems and the emergence of social behavior in such multiagent settings, we were not particularly concerned with providing a detailed social signaling mechanism. It would be interesting to investigate the emergence of richer social behaviors using the same framework, by possibly considering a richer reward space. It may also be interesting to further explore the relation between our work and evolutionary game theory, exploring the relation between this discipline and reinforcement learning [24].

References

[1] E. Deci and R. Ryan, *Intrinsic Motivation and Self-Determination in Human Behavior.* Plenum Press, 1985.

[2] F. Kaplan and P. Oudeyer, "Intrinsically motivated machines," in *50 Years of Artificial Intelligence*, M. Lungarella, F. Iida, J. Bongard, and R. Pfeifer, Eds., 2007, pp. 304–315.

[3] P. Oudeyer, F. Kaplan, and V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE Trans. Evolutionary Computation*, vol. 11, no. 2, pp. 265–286, 2007.

[4] J. Schmidhuber, "Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990-2010)," *IEEE Trans. Autonomous Mental Development*, vol. 2, no. 3, pp. 230–247, 2010.

[5] A. Barto and O. Şimşek, "Intrinsic motivation for reinforcement learning systems," in *Proc. 13th Yale Workshop on Adaptive and Learning Systems*, 2005, pp. 113–118.

[6] A. Stout, G. Konidaris, and A. Barto, "Intrinsically motivated reinforcement learning: A promising framework for developmental robot learning," in *Proc. AAAI Symp. Developmental Robotics*, 2005.

[7] M. Schembri, M. Mirolli, and G. Baldassarre, "Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot," in *Proc. 6th Int. Conf. Development and Learning*, 2007, pp. 282–287.

[8] F. Kaplan and P. Oudeyer, "Motivational principles for visual know-how development," in *Proc. 3rd Int. Workshop on Epigenetic Robotics*, 2003, pp. 73–80.

[9] G. Konidaris and A. Barto, "An adaptive robot motivational system," in *Proc. 9th Int. Conf. Simulation of Adaptive Behavior*, 2006, pp. 346–356.

[10] O. Şimşek and A. Barto, "An intrinsic reward mechanism for efficient exploration," in *Proc. 23rd Int. Conf. Machine Learning*, 2006, pp. 833–840.

[11] S. Singh, R. Lewis, and A. Barto, "Where do rewards come from?" in *Proc. 31st Annual Conf. Cognitive Science Society*, 2009, pp. 2601–2606.

[12] 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.

[13] J. Sorg, S. Singh, and R. Lewis, "Internal rewards mitigate agent boundedness," in *Proc. 27th Int. Conf. Machine Learning*, 2010, pp. 1007–1014.

[14] A. Ng, D. Harada, and S. Russel, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Proc. 16th Int. Conf. Machine Learning*, 1999, pp. 278–87.

[15] J. Maynard-Smith, *Evolution and the Theory of Games.* Cambridge University Press, 1982.

[16] T. Bergstrom, "On the evolution of altruistic ethical rules for siblings," *American Economic Review*, vol. 85, no. 1, pp. 58–81, 1995.

[17] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction.* MIT Press, 1998.

[18] A. Moore and C. Atkeson, "Prioritized sweeping: Reinforcement learning with less data and less real-time," *Machine Learning*, vol. 13, pp. 103–130, 1993.

[19] S. Niekum, A. Barto, and L. Spector, "Genetic programming for reward function search," *IEEE Trans. Autonomous Mental Development*, vol. 2, no. 2, pp. 83–90, 2010.

[20] 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.

[21] D. Dörner, *Bauplan für eine Seele [Blueprint for a Soul].* Reinbeck: Rowohlt, 1999.

[22] J. Bach, *Principles of synthetic intelligence: PSI, an architecture of motivated cognition.* Oxford University Press, 2009.

[23] F. de Waal, "Putting the altruism back into altruism: The evolution of empathy," *Annual Rev. Psychology*, vol. 59, no. 1, pp. 279–300, 2008.

[24] K. Tuyls and A. Nowé, "Evolutionary game theory and multi-agent reinforcement learning," *Knowledge Engineering Review*, vol. 20, pp. 63–90, 2005.