Reward-Sharing Relational Networks in Multi-Agent Reinforcement Learning as a Framework for Emergent Behavior

Hossein Haeri
Dept. of Mechanical Engineering
University of Massachusetts Lowell
hossein_haeri@uml.edu

Reza Ahmadzadeh
Dept. of Computer Science
University of Massachusetts Lowell
reza_ahmadzadeh@uml.edu

Kshitij Jerath
Dept. of Mechanical Engineering
University of Massachusetts Lowell
kshitij_jerath@uml.edu

ABSTRACT

In this work, we integrate 'social' interactions into the MARL setup through a user-defined relational network and examine the effects of agent-agent relations on the rise of emergent behaviors. Leveraging insights from sociology and neuroscience, our proposed framework models agent relationships using the notion of Reward-Sharing Relational Networks (RSRN), where network edge weights act as a measure of how much one agent is invested in the success of (or 'cares about') another. We construct relational rewards as a function of the RSRN interaction weights to collectively train the multi-agent system via a multi-agent reinforcement learning algorithm. The performance of the system is tested for a 3-agent scenario with different relational network structures (e.g., self-interested, communitarian, and authoritarian networks). Our results indicate that reward-sharing relational networks can significantly influence learned behaviors. We posit that RSRN can act as a framework where different relational networks produce distinct emergent behaviors, often analogous to the intuited sociological understanding of such networks.
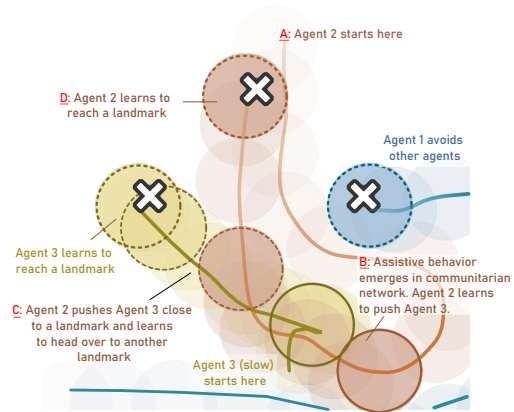
KEYWORDS

Multi-agent Systems, Reinforcement Learning, Social Simulation

1 INTRODUCTION

Recent works in the field of Multi-Agent Reinforcement Learning (MARL) are taking the first steps towards developing a better understanding of interactions between artificially intelligent (AI) agents and their resulting emergent behaviors. While several interesting advances have been made thus far [10, 17, 20, 33], we still lack a broader framework that formulates the MARL problem in a manner that could subsequently generate a theory of emergent behaviors for a network of interacting AI agents. Such a theory of emergent behavior will require (a) a large number of AI agents with (b) complex interactions between the agents that subsequently produce (c) behaviors or dynamics that are discernible only at a global scale [12, 19, 30, 32]. Within the context of MARL problems, this implies the need for a framework that can encapsulate complex agent interactions, learn associated policies, and measure global-scale dynamics (i.e. emergent behavior).

Figure 1: Test episode for a trained communitarian (fully-connected) relational network system. Three agents try to reach three different landmarks (crosses), while agent 3's (yellow) movement is systematically hindered (slowed down). As expected of a communitarian social system, we observed the emergence of an assistive behavior where agent 2 (red) learns to push agent 3 towards another landmark before reaching its own landmark.

Among the several ongoing MARL-related efforts in the research community, we are interested in approaches where learning happens not only as a result of interaction with the environment but also through complex 'social' interactions among agents. To be more specific, Training agents using a decentralized method [4, 25] enables agents to co-evolve, therefore, learning becomes a *collective* process. Moreover, while many state-of-the-art approaches primarily focus on either cooperative or competitive relationships, as illustrated in Figure 2 (a) and (b), the designed MARL framework must go beyond and allow the learned behaviors to span an entire continuous spectrum of social behaviors, as illustrated in Figure 2 (c). In many cases, such behaviors may only be possible via the implementation of distinct reward structures for disparate agents, as presented in Sections 3 and 5.

In this paper, we leverage findings from sociology and neuroscience to create a novel relational learning framework for MARL that can incorporate complex social interactions, which potentially increases our ability to scale and generate emergent behaviors in AI 'societies' or 'ecosystems'. We seek to demonstrate that this framework can produce an effective, systematic approach for implementing more complex relational network structures that address the needs of the MARL research community. There has been limited

prior work on the role of *relational* networks in collective learning in multi-agent systems [28]. This is mainly because, though very common, complex networks arise in environments with more than a few agents. A consequence of working with several agents is that the state-action space suffers from the well-known 'curse of dimensionality', which many existing policy optimization methods are unable to cope with.

We also present the concept of Reward-Sharing Relational Networks (RSRN), implemented into the framework of Multi-Agent Reinforcement Learning (MARL) systems. Our RSRN-MARL framework helps determine how much each agent is invested in the success of (or 'cares' about) other agents, quantified through their individual and relational reward signals. More importantly, in the long term, the notion of reward-sharing relational networks provides a framework and building blocks for creating and evaluating the performance of emergent behaviors in agent 'societies'. We evaluate the effects of reward-sharing relational networks via two different reward sharing implementations and for six different network configurations. Each relational network is examined in a 3-agent MARL environment and analogies to sociological structures (such as self-interested, tribal, authoritarian, and communitarian networks) are provided.

2 RELATED WORK

Reward sharing in reinforcement learning is not in itself a new concept, with prior works on the topic dating back to the early 2000s [7, 21, 23]. One of the first instances of reward sharing in reinforcement learning discussed the idea using the notion of direct rewards (received as a result of an agent's own actions), and indirect rewards (received by other agents because of an agent's action) [23]. We use similar notions in later sections where we refer to individual (or direct) rewards and relational (or indirect) rewards in a network context. However, most related work in the domain has focused on game-theoretic approaches, which inherently limits the discussion of the problem to either a competitive [15] or collaborative context [18, 26]. Other recent works have sought to move away from this dichotomous approach using the notion of mixed-motive games, even introducing scalar representations of prosocial propensity (such as social value orientation) [22]. However, a broader framework that eschews the dichotomous competitive vs. collaborative mindset for a more general relational network approach has largely been missing [28]. Unlike these prior studies, the presented work borrows from the fields of sociology and neuroscience to provide a framework for emergent behavior by implementing reward sharing using an explicitly-defined 'social' or relational network.

There are also other recent efforts in the field of MARL that have studied multi-agent systems whose agents are able (or have the potential) to communicate with each other via a special communication channel to share information and coordinate with each other. [10, 14, 16, 34]. These works have often been grouped collectively into the category of 'networked' MARL, which may, at first glance, appear to have some similarity to the presented work. However, in these related works, agents usually learn the communication skills using a single team-level global reward function [34]. Significant research efforts in this domain are directed towards solving, among others, the problem of communication-constrained distributed reinforcement learning.
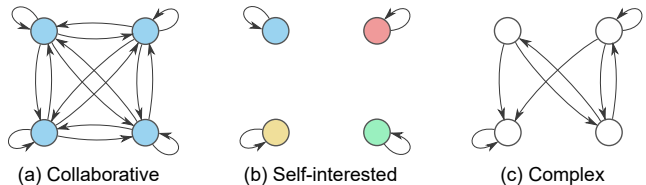
Figure 2: In a decentralized MARL environment, agents co-evolve and learn simultaneously. (a) Collaborative agents (b) Self-interested agents (c) Generalized complex network structure. Prior works have largely focused on problems associated with (a) and (b).

Our study, however, is focused on the high-level 'relationships' between the agents. More specifically, the presented work focuses on capturing agent interactions in a 'relational' graph that is also leveraged for reward-sharing and collective agent learning in the MARL framework. The weights of the graph edges between nodes (i.e. agents) represent how valuable another agent's success is to any given agent. This notion of reward sharing in a relational network is only tangentially related to the communication-focused research efforts performed under the umbrella of networked MARL. Section 4 will provide additional details on this topic, with a discussion on different network structures and their impact on team rewards.

3 METHODOLOGY

The primary innovation of the presented work is in the integration of relational networks with existing MARL methods which results in a novel framework that helps us understand how 'social' interactions between agents generate emergent behaviors. These relational connections help determine how rewards obtained by one agent are shared with and drive the actions of another agent, and reflect which entities the agent 'cares about'. Additional details on the sociological inspiration associated with this notion are provided in Section 4. As Figure 3 shows, our proposed framework adds a reward sharing block, called the Reward-Sharing Relational Network (RSRN), to a typical MARL flow diagram. The RSRN block takes a vector of individual agent rewards and a user-defined relational network as inputs and produces a vector of relational rewards for each agent as output. The output of shared relational rewards is used by the policy optimizer to generate a set of actions for every agent in the multi-agent environment.

3.1 Problem Formulation

We define a Relationally Networked Decentralized Partially Observable Markov Decision Process (RN-Dec-POMDP) as a tuple $(\mathcal{S}, \mathcal{A}, O, \mathcal{T}, \mathcal{R}, \mathcal{G})$, where $\mathcal{S} = \{S_1, S_2, ..., S_N\}$, $\mathcal{A} = \{A_1, A_2, ..., A_N\}$, $O = \{O_1, O_2, ..., O_N\}$, and $\mathcal{R} = \{R_1, R_2, ..., R_N\}$ are the joint set of individual states, actions, observations, and rewards, respectively. The tuple element $\mathcal{G} = (V_{\mathcal{G}}, E_{\mathcal{G}}, W_{\mathcal{G}})$ represents the ordered collection of all agents as vertices in the set $V_{\mathcal{G}}$, all binary agent relations as directed edges in the set $E_{\mathcal{G}}$, and edge weights as $W_{\mathcal{G}}$. With reference to Figure 4, the presence of an edge between two agents represents that they are related. The direction of the edge from a first agent to a second agent represents that the actions of the first agent are driven by the rewards obtained by the second. This may
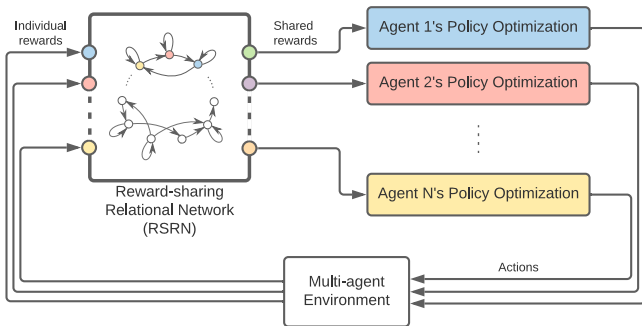
Figure 3: Agents obtain individual rewards from the multi-agent environment. Shared rewards are evaluated using the Reward-Sharing Relational Network (RSRN) and depend on agent-agent relations. Different network structures generate different shared rewards, which are then used by the policy optimizer(s) and produce distinct emergent behaviors.
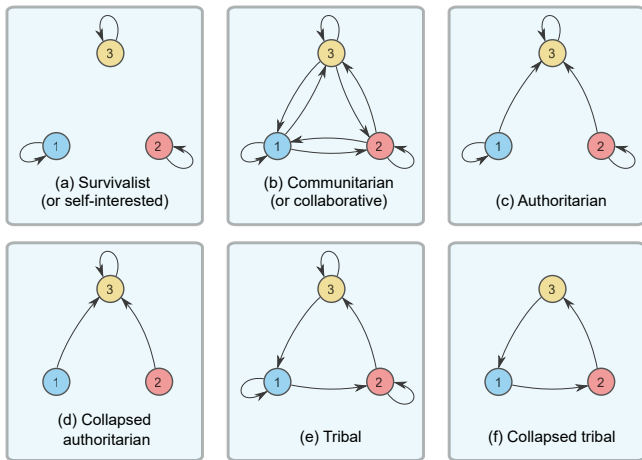
Figure 4: Examples of sociology-inspired relational network structures. Each of these relational networks is used as an RSRN in the 3-agent MARL system, and the resulting emergent behaviors are examined in Section 5. An agent's actions may be governed by the rewards obtained by more than one agent (including itself). An arrow directed from a first agent to a second represents that the first agent's actions are governed by the rewards obtained by the second agent.

be understood as the second agent *sharing rewards* with the first agent, or that the first agent 'cares about' the success of the related second agent. Consequently, the first agent is likely to learn policies that benefit the other agent. Of course, the relational network also allows for self-directed edges, so an agent's actions can also be driven by its own rewards. The nature of these relationships is encapsulated in the matrix $W_{\mathcal{G}}$ which denotes the weights associated with the edges.

At each time step t, each agent $i \in \{1, 2, ..., N\}$, first perceives the environment in the form of an observation $o_i \in O_i$, where O_i is a set of its individual observations. Agent i then takes action $a_i \in A_i$ based on its policy (in this case, deterministic) $\pi_i : O_i \to A_i$ and transitions (along with other agents) from to the next state according to the joint probabilistic transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$. Once an agent reaches its next state, it receives a reward r_i according to its reward function $R_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

The relational network weights $W_{\mathcal{G}}$ can be expressed as a matrix whose elements $w_{i,j} \in \mathbb{R}$ indicate how much agent i 'cares about' the success of agent j, based on the individual reward obtained by agent j. Thus, actions of agent i may be driven in part by the individual rewards of agent j, if these two agents are related, i.e. if $w_{i,j} \neq 0$. For example, a society with only self-interested agents, who are unrelated to and do not 'care about' other agents, can be represented by the identity matrix. On the other hand, a communitarian society, that is highly collaborative and where relations exist between all agents, can be expressed by the all-ones matrix. Figure 4 illustrates these two networks as well as other examples of RSRNs which will be discussed in section 4. In general, due to the fact that the relational network includes directed edges, the weight matrix $W_{\mathcal{G}}$ is not necessarily symmetric. For example, in the tribal relational network shown in Figure 4(e), $w_{1,2} = 1$ (i.e. Agent 1 is invested in the success of Agent 2), but $w_{2,1} = 0$ (i.e. Agent 2 does not 'care' about Agent 1).

Integrating relational networks into the Dec-POMDP framework expands the research avenues associated with MARL to a spectrum of possible policies and behaviors depending on the definition of the weight matrix. Since most typical social systems are sparse, many weights $w_{i,j}$ are zero (e.g. small-world networks) [2]. This key insight enables us to limit the reward sharing of an individual to only a few other agents. It also plays an important role in the tractability of finding the optimal policy, and perhaps an even greater role in future works seeking to scale the problem to larger agent populations.

3.2 Scalarization of Individual Rewards

In our RSRN-MARL framework, agents try to maximize their long-term shared return $\bar{R}_i$ by accumulating the discounted shared relational rewards $\bar{r}_i$ across a finite horizon T as follows:

$$\bar{R}_i = \mathbb{E}\left(\sum_{k=0}^{T} \gamma^k \bar{r}_{i,k}\right) = \mathbb{E}\left(\sum_{k=0}^{T} \gamma^k f(\mathbf{r}_k, \mathbf{w}_i)\right) \tag{1}$$

where $\bar{r}_i$ represents the shared relational reward of agent i, which incorporates the individual rewards of all related agents (and itself). Specifically, the scalarization function $f(\mathbf{r}_k, \mathbf{w}_i)$ maps all individual rewards $\mathbf{r}_k = [r_1, r_2, ..., r_N]_k^\top$ at time step k to a single shared relational reward value $\bar{r}_{i,k}$ for agent i according to the agent-specific relational weight vector $\mathbf{w}_i = [w_{i,1}, w_{i,2}, ..., w_{i,N}]$, which determines how agent i is related to other agents.

It is possible to choose the scalarization function from several different alternatives, with the most commonly used choice often being a simple weighted sum across all the individual rewards, given by $\bar{r}_i = f_s(\mathbf{r}, \mathbf{w}_i) = \sum_{j=1}^{N} w_{i,j} r_j$. However, we found that the Weighted Product Model (WPM), given by $\bar{r}_i = f_p(\mathbf{r}, \mathbf{w}_i) = \prod_{j=1}^{N} r_j^{w_{i,j}}$, performed much better in our experiments while preserving the intuitive, qualitative relational between agents. Moreover, assuming that the rewards are positive, both scalarization functions are strictly increasing for positive weights and are strictly decreasing

for negative weights. Hence, they both satisfy the minimal assumption of scalarization functions [31], i.e. the higher reward values map to higher scalarized values. In this study, we used WPM scalarization due to the fact that it dramatically lowers the shared reward when even one of the individual rewards is close to zero. Therefore, trainers are strictly promoted to care equally about all the agents who are related with equal weights.

3.3 Policy Optimization

The presented work on relational networks does not rely on any specific policy optimization method. Depending on the application, we can use centralized or decentralized trainers. In centralized training, a single policy is optimized using all agents' action-observation feedback [33]. Alternatively, one can use a set of single-agent policy optimizers to cope with this challenge. Although decentralized approaches are more general [11, 27], they have to deal with the non-stationarity and may not converge. Here, we briefly discuss one of these algorithms, Multi-agent Deep Deterministic Policy Gradient (MADDPG) [20] that is an actor-critic method based on Deep Q-Networks (DQN) [24]. In this algorithm, agent policies $\pi = \{\pi_1, ..., \pi_N\}$ are represented using neural networks parameterized by $\theta = \{\theta_1, ..., \theta_N\}$. To optimize the policy, the gradient of the expected return for each agent can be written as $\nabla_{\theta_i} J(\theta_i) = \mathbb{E}(\nabla_{\theta_i} \log \pi_i(a_i|o_i) Q_i^\pi)$ [20]. If we consider continuous deterministic policies μ_i parameterized by θ_i, the gradient can be written as:

$$\nabla_{\theta_i} J(\mu_i) = \mathbb{E}(\nabla_{\theta_i} \mu_i(a_i|o_i) \nabla_{a_i} Q_i^\mu) \tag{2}$$

where $Q_i^\pi(\boldsymbol{o}, \boldsymbol{a})$ is the centralized action-value function, $\boldsymbol{o} = \{o_1, ..., o_N\}$ is the set of multi-agent observations, and $\boldsymbol{a} = \{a_1, ..., a_N\}$ represents the set of multi-agent actions. The action-value function $Q_i^\pi(\boldsymbol{o}, \boldsymbol{a})$ is then updated to minimize the loss function:

$$\mathcal{L}(\theta_i) = \mathbb{E}\left[\left(Q_i^\pi - (\bar{r}_i + \gamma Q_i^{\pi'})\right)^2\right] \tag{3}$$

where π' is a set of target policies with delayed parameters θ_i'. Note that unlike MADDPG that uses the individual rewards r_i, we calculate the loss using agents' shared relational rewards $\bar{r}_i$.

4 STRUCTURE OF REWARD-SHARING RELATIONAL NETWORKS

In this section, we will discuss how our work on reward-sharing relational networks (RSRN) is inspired in part by existing work on sociology [5, 6] and neuroscience [9, 13]. To form a framework for emergent behavior, we need to better understand how relational networks should be structured. We begin with two insights into the nature of learning in social systems comprised of primates (including humans) or other mammals [8]. The first insight is that individual learning occurs in social settings [1]. Specifically, it has been shown that learning amongst primates and other mammals is governed by whom the individuals can relate to. Individual learning is driven by actions ranging from simple mimicking to complex evaluative behaviors, all within the presence of other *socially-related* individuals [29]. The presented work on reward-sharing relational networks formalizes this notion within the context of MARL.

The second insight borrows from neuroscience and indicates that learning individuals can connect or relate to a *limited* number

of their counterparts [9]. This finding originates from studies of the size of the neocortex in primate brains and its relationship to the group size of the social systems these primates inhabit. Building on these two insights, we propose that reward-sharing and learning in a multi-agent system should: (a) be governed by whom an agent is related to, and (b) be limited to a small number of relations to obtain different types of emergent behaviors. The relational networks thus generated are used for training the agents, and may potentially generate distinct emergent behaviors depending on the structure of the relational network. We evaluate the proposed framework via simulations of a 3-agent system in different environments and the results are presented in Section 5.

4.1 Implications of Relational Network Structures

The structure of relational networks that exist in social systems (such as the human society) have deep implications on the policies learned and emergent behaviors observed in such systems [1, 6]. For example, strictly hierarchical relational networks produce authoritarian emergent behaviors, where individuals are significantly driven by the states, actions, and performance of the individuals higher up in the hierarchy [3]. One may intuitively expect to observe similar emergent behaviors in other similarly structured 'social' system of AI agents, such as in our RSRN-MARL framework.

To provide the reader with a better description, we now discuss other examples of relational networks with reference to Figure 4. As seen in Figure 4(a), *survivalist* relational networks consist of individuals that are primarily self-interested and focused on self-preservation. In a MARL system, this may translate into keeping one's reward for oneself, so that an agent's actions are governed only by the reward it gets for itself. The associated relational network weight matrix W_G is, in its simplest form, an identity matrix, with no relationships with other individuals. Prior works may model this behavior via non-cooperative games. On the contrary, *communitarian* relational networks are highly collaborative in nature, and each individual is related to every other individual and 'cares' about them, as shown in Figure 4(b). In a MARL system, this may be implemented as reward sharing with all agents, such that the policies learned by an agent are a function of the rewards of *all* other agents, and not just its own rewards. Prior works may model this behavior to some extent as collaborative games [20].

To the best of our knowledge, the other relational networks do not have distinct, formalized counterparts in prior MARL literature, and several other network configurations not shown may also be evaluated. Figure 4(c) represents an *authoritarian* relational network which is defined by its hierarchical elements [6]. Specifically, in the authoritarian relational network, all individuals in the social system are invested in their own success, but their actions are also driven by the need to ensure the success of those above them in the hierarchy. E.g., in the context of a 3-agent MARL problem shown in Figure 4(c), the directed edge from agent 1 to the 'higher' agent 3 indicates that agent 1's actions are governed by agent 3's rewards. However, there is no directed edge from agent 3 to the 'lower' agent 1, as agent 3's policies are not driven by the individual rewards of agent 1. A devolved version of the network in Figure 4(c) may represent a *collapsed authoritarian* relational network (Figure 4(d)),

where the hierarchy is maintained, but exists solely for the purpose of benefiting the individuals higher in the hierarchy.

Similarly, Figure 4(e) represents a *tribal* relational network, where along with being driven by one's own success, an individual is also driven by the success of specific other individuals with whom it shares a familial relationship. In the MARL context, and especially in the 3-agent network, we represent this as directed edges from one agent to the next. These tribal relational networks will be easier to visualize and comprehend with larger multi-agent systems. A devolved version of the network in Figure 4(e) may represent a *collapsed* or *poorly scaffolded tribal* relational network as shown in Figure 4(f). In this relational network, individuals only care about the success of their familial relations and are not concerned with their own success.

5 SIMULATION AND RESULTS

In this section, we implement an RSRN on a 3-agent MARL environment and evaluate the performance of the agents under different network structures. The environment is simulated using the Multi-agent Particle Environment (MPE) [20] where agents are simulated as physical elastic objects. Both state and action spaces are continuous and agents have been trained using our framework by integrating the Relational Network and Multi-Agent Deep Deterministic Policy Gradient (MADDPG) policy optimization algorithm [20]. We used 2 fully-connected 64-unite layers for policy networks which are optimized with the learning rate of 0.01, batch-size of 2048, and the discount ratio of 0.99.

As shown in Figure 1, We consider the scenario where three agents try to reach three unlabeled landmarks, i.e. they get rewarded if they reach any landmark. To make the multi-agent environment more complex and create opportunities to observe emergent behaviors arise, we limit the mobility of one of the agents. Specifically, agent 3 is hindered systematically so it cannot move as fast as other agents, and may be unable to reach a landmark within a single episode.

To evaluate the collective, emergent behavior of the agents, we examine 6 different relational network configurations as shown in Figure 4. The individual performance of each agent i is measured through its individual reward r_i, which is $e^{-d_i^2/\sigma^2}$ if d_i is less than 0.2 and is zero otherwise, where d_i denotes the Euclidean distance to the closest landmark and constant σ^2 is set to 0.1. It should be noted that the landmark locations are fixed in these simulations but agents are initialized at random positions.

After training agents, we run test episodes with the learned policies to calculate their mean individual and shared relational rewards which correspond to their individual and social performances, respectively. Figure 5(a) shows individual rewards which determine how successful each agent reaches the landmarks. Similarly, Figure 5(b) shows shared relational rewards which indicate the performance of the related agents. Additional discussion related to this topic can be found in Section 5.1. These figures and associated videos[1] help reveal insights into how different relational networks produce distinct emergent behaviors, as well as the performance of individual agents.

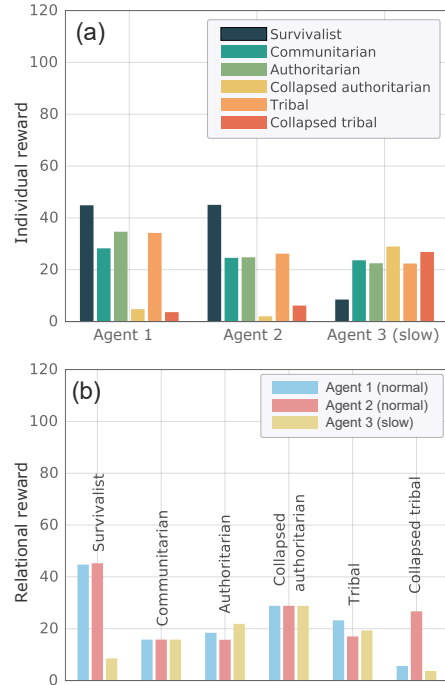[1]Videos are available at sites.google.com/view/marl-rsrn

Figure 5: (a) Individual rewards and (b) relational rewards using WPM scalarization. Here, all rewards are averaged across 5000 test episodes, after training over 500,000 episodes.

5.1 Examining Emergent Behaviors in Relational Networks

In the following paragraphs, we will briefly discuss the emergent behaviors for each network configuration.

Survivalist (or self-interested): In this network configuration, the shared relational reward for each agent is the same as its own individual reward, since no relations exist between agents. Agents 1 and 2 easily learn policies that generate significant individual rewards. However, the constrained agent 3 is unable to collect meaningful rewards. In fact, as is evident from Figure 6(a), agent 3 even loses some of its initial randomly collected reward because the other agents learn to serve their own interest and push agent 3 away from a closer landmark if needed. As shown in the tests results in 5(a), agent 3 fares much poorly than the other agents in this survivalist relational network.

Communitarian (or collaborative): In this setup, $w_{i,j} = 1, \forall i, j$. Consequently, all agents receive identical rewards, and they form a homogeneous team where no one has any reward privileges over others. Figure 5(a) indicates that all three agents receive similar individual rewards in the test phase. More importantly, this is a direct result of the identical shared relational rewards received by the three agents, as shown in Figure 5(b). While agents 1 and 2 receive lower rewards as compared to the survivalist relational network, each agent in a communitarian network is taken care of.

Authoritarian: In this relational network (RN), agents 1 and 2 weigh their own rewards as much as they weigh the rewards received by the slow agent 3 ('the monarch'). The emergent behavior observed
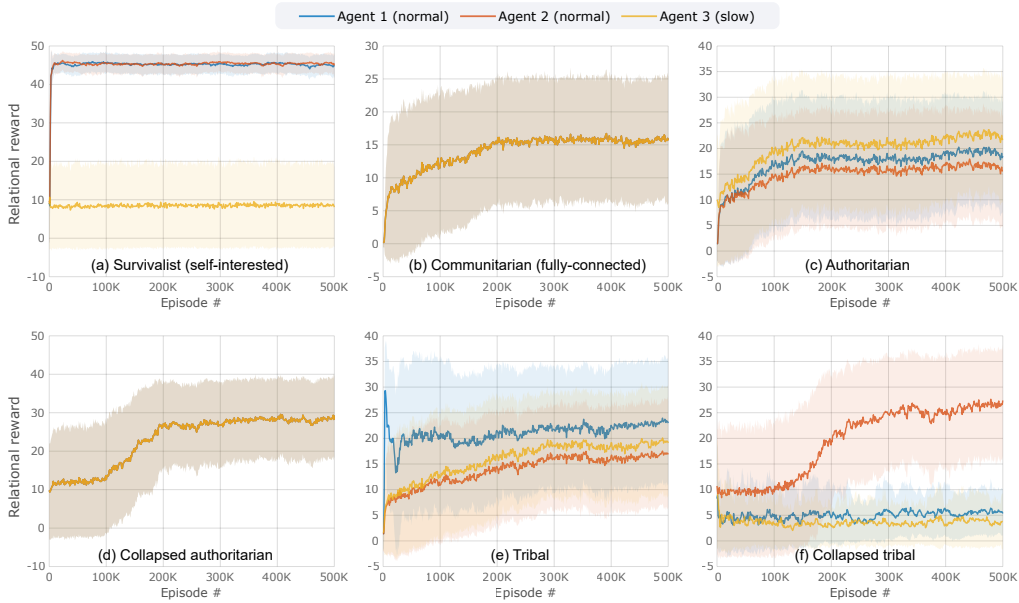
Figure 6: The relational reward of agents as they are being co-trained with different RSRNs.

here includes agents 1 and/or 2 assisting agent 3 to the landmark before reaching other landmarks themselves.

Collapsed Authoritarian: This relational network demonstrates even more interesting emergent behaviors than the authoritarian RN. As a potentially devolved version of the authoritarian network, this RN demonstrates behaviors associated with extreme subservience to the rewards of agent 3. In other words, agents 1 and 2 learn policies that only benefit agent 3 ('the dictator'). These agents receive minimal individual rewards for themselves, since $w_{1,1} = w_{2,2} = 0$ (Figure 5(a)), but significant relational rewards, since $w_{1,3} = w_{2,3} = w_{3,3} = 1$ (Figure 5(b)). In video evidence, agents 1 and 2 are seen to assist agent 3 to the landmark and either continue to loiter around agent 3 to make minor adjustments that may increase its reward, or eventually wander away in the environment.

Tribal: Each agent i in this configuration cares equally about itself and a specific 'friend' or familial agent j, though this is not reciprocated by agent j. In the specific configuration considered in 4(e), the relationship is cyclic, so each agent cares about the next. This produces a sustainable tribal relationship, as evinced from the comparable individual and relational rewards across all three agents seen for tribal RN in Figure 5.

Collapsed Tribal: In this relational network, we observe a devolved tribal or familial relationship, where individual agents no longer learn policies associated with their own rewards but only of their related agent's rewards. Agent 1 performs poorly with low individual and relational rewards, since only agent 3 'cares' about it, and agent 3 is systematically hindered or slow. Hence, agent 3 is unable to learn requisite actions that would increase the rewards of agent 1. Agent 2, on the other hand, can move fast and learn optimal policies that increase individual rewards for agent 3, and consequently the shared relational rewards for agent 2. Thus, we observe from Figure 5 that agent 3's individual reward and agent

2's relational reward are significant, respectively. Moreover, agent 3's individual rewards are higher than those received in a tribal RN, but this is at the cost of severely reduced performance of the other agents. Video evidence indicates that agent 2 hovers close to agent 3, even after the latter has reached a landmark. While it was intuitively expected that agent 1 would intervene and push agent 2 towards another landmark after agent 3 had been assisted, we did not observe such behavior. One possible reason could be that agents 1 and 2 are equally fast, and agent 1 is unable to learn to push agent 2 in any meaningful way.

6 CONCLUSIONS AND FUTURE WORK

Driven by insights from sociology and neuroscience, we have proposed a Reward-Sharing Relational Network (RSRN) multi-agent reinforcement learning framework that formalizes social learning by determining how agents 'care about' each other. The agents learn policies that maximize their shared relational rewards, which are constructed according to the individual rewards of all related agents. This framework has the potential to act as a foundation for studying agent interactions and the generation of emergent behaviors in societies or ecosystems of learning agents.

We have demonstrated the novelty and effectiveness of this approach using a 3-agent MARL system with various relational network configurations, where agents learn to reach fixed landmarks. As Figure 5(a) showed, different relational networks result in distinct learned behaviors of finding landmarks and individual agent performances. The learned policies often include emergent behaviors not explicitly defined in the problem formulation. E.g., agents in a *survivalist* relational networks quickly learned to go to a landmark, even pushing slower agents out of the way, though no such behavior was included in the problem formulation. On the other hand, agents trained with a *communitarian* relational

network adapt and learn to distribute their tasks according to their capabilities. The system exhibits emergent behavior as fast-moving agents learn to assist the slow-moving agent towards the landmark, before finding a landmark of their own. Similar emergent behaviors manifest in *authoritarian* and *tribal* relational networks as well. E.g., in authoritarian relational networks, fast-moving agents lower in the hierarchy assist slow-moving agents higher in the hierarchy. However, these services are not afforded to slow-moving agents if they exist lower in the authoritarian hierarchy. Collapsed authoritarian and tribal networks also demonstrate some extreme behaviors where some agents are subservient to the rewards received by other agents and continue to hover around slow-moving agents, even after they have been assisted towards a landmark.

Among the several challenges associated with studying emergent behaviors in the MARL problems, a key issue holding the field back is the apparent intractability and difficulty in scaling the problem to a large number of agents, which happens to be a key requirement to study emergent dynamics of interacting AI agents. Our future works will be directed towards leveraging the Reward-Sharing Relational Networks in Multi-agent Reinforcement Learning problems to specifically tackle the challenge of scalability. Potential approaches to expand upon the presented work include (a) the expansion of RSRNs to dynamic or evolving relational networks (unlike the current work, which assumed static network configurations), (b) the use of fractional or negative weights to model less-than-collaborative or hostile relationships, respectively, and (c) estimating the rewards obtained by related individuals. We expect these advances to provide the first steps towards building a theory of emergent behaviors in interacting AI agents.

ACKNOWLEDGMENTS

REFERENCES

[1] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. 2011. Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing* 7, 6 (2011), 643–659.

[2] Luis A Nunes Amaral, Antonio Scala, Marc Barthelemy, and H Eugene Stanley. 2000. Classes of small-world networks. *Proceedings of the national academy of sciences* 97, 21 (2000), 11149–11152.

[3] Joel Aronoff and Lawrence A Messe. 1971. Motivational determinants of small-group structure. *Journal of Personality and Social Psychology* 17, 3 (1971), 319.

[4] Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. 2019. Emergent tool use from multi-agent autocurricula. arXiv. *arXiv preprint arXiv:1909.07528* (2019).

[5] Robert F. Bales. 1950. A Set of Categories for the Analysis of Small Group Interaction. *American Sociological Review* 15, 2 (apr 1950), 257. https://doi.org/10.2307/2086790

[6] Don Edward Beck and Christopher C Cowan. 2014. *Spiral dynamics: Mastering values, leadership and change*. John Wiley & Sons.

[7] Yoshinobu Bochi, Tadachika Ozono, and Toramatsu Shintani. 2003. A direct-indirect reward sharing model in multiagent reinforcement learning. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*. 940–941.

[8] Frans De Waal. 2010. *The age of empathy: Nature's lessons for a kinder society*. Broadway Books.

[9] R. I.M. Dunbar. 1992. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution* 22, 6 (jun 1992), 469–493. https://doi.org/10.1016/0047-2484(92)90081-J

[10] Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in neural information processing systems*. 2137–2145.

[11] Jakob Foerster, Nantas Nardelli, Gregory Farquhar, Triantafyllos Afouras, Philip HS Torr, Pushmeet Kohli, and Shimon Whiteson. 2017. Stabilising experience replay for deep multi-agent reinforcement learning. *arXiv preprint arXiv:1702.08887* (2017).

[12] Hossein Haeri, Kshitij Jerath, and Jacob Leachman. 2020. Thermodynamics-Inspired Macroscopic States of Bounded Swarms. *ASME Letters in Dynamic Systems and Control* 1, 1 (2020).

[13] Philip L Jackson, Andrew N Meltzoff, and Jean Decety. 2005. How do we perceive the pain of others? A window into the neural processes involved in empathy. *Neuroimage* 24, 3 (2005), 771–779.

[14] Ivana Kajić, Eser Aygün, and Doina Precup. 2020. Learning to cooperate: Emergent communication in multi-agent navigation. *arXiv preprint arXiv:2004.01097* (2020).

[15] Erich Kutschinski, Thomas Uthmann, and Daniel Polani. 2003. Learning competitive pricing strategies by multi-agent reinforcement learning. *Journal of Economic Dynamics and Control* 27, 11-12 (2003), 2207–2218.

[16] Angeliki Lazaridou and Marco Baroni. 2020. Emergent Multi-Agent Communication in the Deep Learning Era. *arXiv preprint arXiv:2006.02419* (2020).

[17] Joel Z Leibo, Edward Hughes, Marc Lanctot, and Thore Graepel. 2019. Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. *arXiv preprint arXiv:1903.00742* (2019).

[18] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, Vol. 1. 464–473. arXiv:1702.03037 www.ifaamas.org

[19] Zhengping Li, Cheng Hwee Sim, and Malcolm Yoke Hean Low. 2006. A survey of emergent behavior and its impacts in agent-based systems. In *2006 4th IEEE International Conference on Industrial Informatics*. IEEE, 1295–1300.

[20] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in neural information processing systems*. 6379–6390.

[21] Maja J Matarić. 2001. Learning in behavior-based multi-robot systems: Policies, models, and other agents. *Cognitive Systems Research* 2, 1 (2001), 81–93.

[22] Kevin R McKee, Ian Gemp, Brian McWilliams, Edgar A Duéñez-Guzmán, Edward Hughes, and Joel Z Leibo. 2020. Social Diversity and Social Preferences in Mixed-Motive Reinforcement Learning. *arXiv preprint arXiv:2002.02325* (2020).

[23] Kazuteru Miyazaki and Shigenobu Kobayashi. 2001. Rationality of reward sharing in multi-agent reinforcement learning. *New Generation Computing* 19, 2 (2001), 157–172.

[24] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.

[25] Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P How, and John Vian. 2017. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. *arXiv preprint arXiv:1703.06182* (2017).

[26] Afshin OroojlooyJadid and Davood Hajinezhad. 2019. A review of cooperative multi-agent deep reinforcement learning. *arXiv preprint arXiv:1908.03963* (2019).

[27] Georgios Papoudakis, Filippos Christianos, Arrasy Rahman, and Stefano V Albrecht. 2019. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737* (2019).

[28] Alexander Peysakhovich and Adam Lerer. 2018. Prosocial learning agents solve generalized stag hunts better than selfish ones. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, Vol. 3. 2043–2044. arXiv:1709.02865

[29] Charles Pezeshki and Ryan Kelley. 2014. It's all about relationships: how human relational development and social structures dictate product design in the socio-sphere. In *Product Development in the Socio-sphere*. Springer, 143–168.

[30] Ilya Prigogine and Isabelle Stengers. 2018. *Order out of chaos: Man's new dialogue with nature*. Verso Books.

[31] Roxana Rădulescu, Patrick Mannion, Diederik M Roijers, and Ann Nowé. 2020. Multi-objective multi-agent decision making: a utility-based analysis and survey. *Autonomous Agents and Multi-Agent Systems* 34, 1 (2020), 10.

[32] Claudia Szabo and Yong Meng Teo. 2015. Formalization of weak emergence in multiagent systems. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 26, 1 (2015), 1–25.

[33] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[34] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. 2019. Decentralized Multi-Agent Reinforcement Learning with Networked Agents: Recent Advances. *arXiv preprint arXiv:1912.03821* (2019).