Team-based Multi-agent Reinforcement Learning

Luke Liem
Department of Computer Science
UCSD
San Diego, CA 92108
A53231779
lliem@eng.ucsd.edu

Abstract

In multi-agent reinforcement learning (MARL), differentiating between agent intelligence and organization intelligence may hold the key to major breakthroughs.

Most MARL researches focus on improving agent intelligence, then expects these agents to exhibit advanced organization behaviors such as cooperation, alliance and reciprocity. They try to develop complex algorithms that supposedly "outperform" the naive RL algorithm, which learns policies that maximize the individual reward of each agent. These "state-of-the-art" algorithms are mathematically and computationally complex. In many cases, they only work well when agents are given the policy parameters of the other agents [1,4]. This latter requirement is not realistic for many applications and severely limits scalability when the number of agents increases.

This project separates the encoding of agent intelligence from organization intelligence. Agents are programmed with a simple naive algorithm, but they are organized under teams and provided with US versus THEM context. The organization intelligence is separately encoded in the team's culture, which determines how team rewards are doled out to its agents on top of the environmental reward they gather during training.

With the separation of agent and organization intelligences, the methodology becomes mathematically and computationally simple. It can scale easily with the number of agents and teams and it enables teams of agents to achieve a wide range of desired results and behaviors with only slight changes to the team culture and no change to the agents' policy algorithm.

The new approach enables teams of agents to easily exceed the performance of agents trained under "state-of-the art" MARL algorithms. In addition, the use of team reward in culture can lead to agent specialization, which enables a team of specialized agents to build a dominating strategy to a game which is previously intransitive to multiple individual agents.

1. Introduction

Agent vs. Organization intelligence

In the book *Sapiens*, the author Yuval Harari provides evidences that up until 70,000 years ago, *Homo sapiens* was just one of several pre-human species on Earth, after all these pre-human species evolved from a genius of ape 2.5 million years ago. Yet from that moment onward, the *Homo sapiens* catapults itself from the middle of Earth's food chain to the top, in the process wiping out all other pre-human species during the Cognitive Revolution.

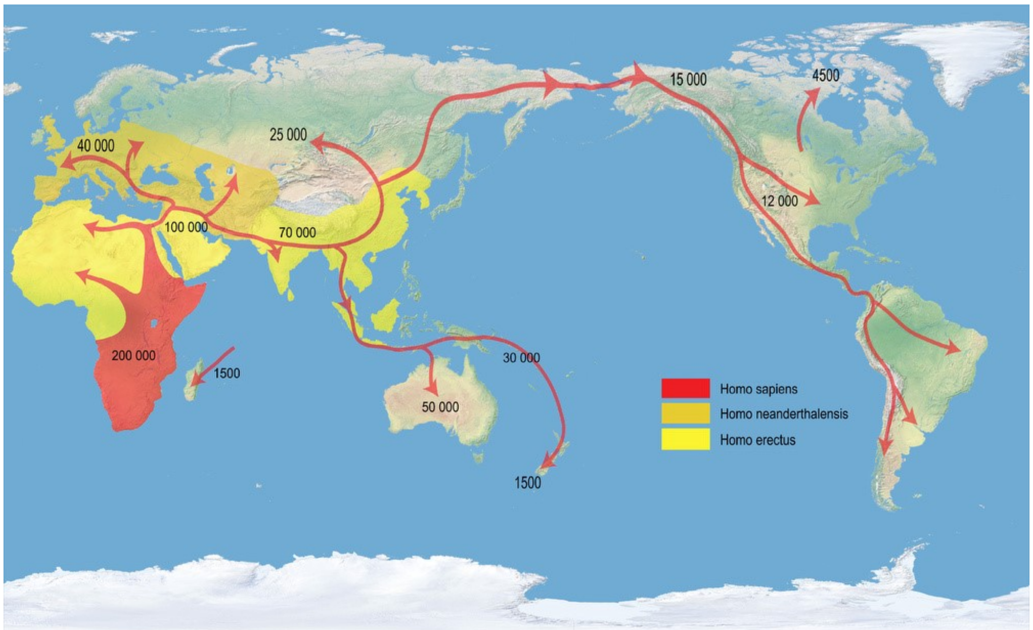


Figure 1: Spread of *Homo sapiens*

In this Revolution, a chance mutation in the *Homo sapiens'* genes gave them the ability to develop "fictive language" – a language that can describe not only concrete objects but also imaginative ones. The fictive language allowed *Homo sapiens* to develop advanced concepts like money, religions, nations and capitalism and enabled them to organize themselves into larger and larger organizations – from hunting packs to tribal communities to nation states, from self-proprietorships to partnerships to multinational corporations.

On an individual basis, the *Homo sapiens* has a smaller brain than the *Homo Neanderthal* and the *Homo Neanderthal* is physically stronger. But since the *Homo sapiens* could organize themselves into hunting and warring parties of over 150 while the *Homo Neanderthals* could only put together parties of 10-20 or so, the former drove the latter into extinction in Europe and the Middle East. The *Homo sapiens* may have less agent intelligence than the *Homo Neanderthals*, but the fictive language gave them superior organization intelligence.

In MARL research, it is important to differentiate between what is agent intelligence which has a biological nature, versus what is organization intelligence which is more cultural. With this distinction in mind, researchers will not fall into the trap of over-designing a *Homo Neanderthal* and then expecting them to come up with organization intelligent concepts such as money, religion, ethics and nation state.

Team-based Agents

In biological world and throughout human history, intelligent agents organize themselves into teams or tribes. The context needed for this to work is simply the agent's ability to distinguish between US and THEM.

Instead of designing super-intelligent agents who may or may not eventually figure out that they need to form tribes, we can design this US versus THEM context into the Environment and the Agent and Team Classes, thus redefining traditional reinforcement learning as an interaction between the Environment and Teams of Agents.

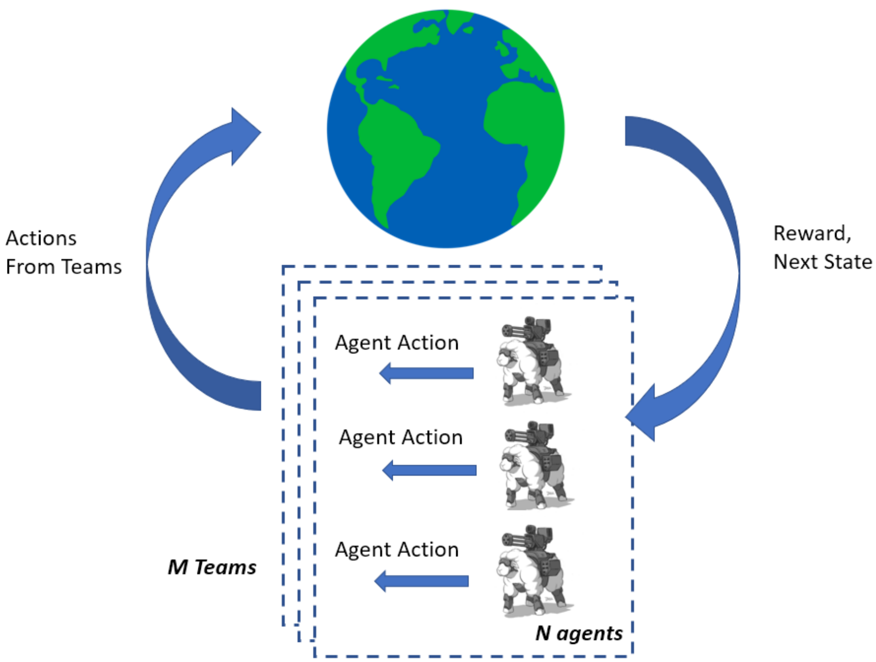


Figure 2: Team-based multi-agent reinforcement learning

Culture and Team Reward

By organizing agents under a team, we separate the encoding of agent intelligence from that of organization intelligence. While the agent's intelligence is programmed within its policy network, we encode the team's organization intelligence within its culture.

Members of a team or a tribe share common culture and behave in specific ways due to intra-team or intra-tribe reward doled out by the team or the tribe. Sometimes, the intra-team reward is material and may be a portion of the environmental reward gathered in the form of livestock and staples. But the most powerful intra-team rewards are non-material. Roman legion's standard-bearer would sometimes throw the legion's eagle into the enemy so that legionnaires threw themselves at the enemy to retrieve it. Greek and Roman generals after a victory were awarded laurel wreaths symbolizing the people's honor and admiration for them. These non-material yet powerful rewards exist only in the imagination of the collective minds of a team or a tribe.

In Team-based MARL, the Team use intra-team reward to shape the agents during training. This reward is doled out based on the agent's specific behavior or on the results achieved. In this manner, agents with simple naive algorithm can be organized into effective teams imbued with superior organization intelligence. Teams of these agents can outperform unorganized teams of more "intelligent" agents the same way *Homo sapiens* out-organized and outperformed *Homo Neanderthals*.

2. Related Works

Many classical researches in MARL test their algorithms on iterated matrix games [2], which pit 2 players against each other with Cooperate/Defect as being the only available agent actions. In 2017, DeepMind proposed that Markov games is a better environment for testing MARL algorithms and introduced sequential social dilemmas (SSD) as a richer framework for MARL research [3]. Even then, most researchers have been focusing their efforts on improving the intelligence of the individual agent through "superior" algorithms, and they claim success when their agents outperform naive agents in games when all agents are "lone wolves" with no social context with each other [1,4].

Our work is unique in (1) our differentiation between agent intelligence and organization intelligence and (2) our redefinition of MARL as an interaction between the Environment and Teams of Agents. Agents are programmed with simple naive algorithm, but are organized under teams and provided with US versus THEM context. The organization intelligence is separately encoded in the team's culture, which determines how team rewards are doled out to its agents on top of the environmental reward they gather during training. We believe our work is unique in proposing the use of this "imaginary" team reward to shape the learning of teams of agents during training, but not in game play.

We believe ours is also the first paper that provides analysis about (1) how a team of agents can device a dominating strategy for a game which is previously intransitive to multiple individual agents, and (2) how dominating strategy in MARL game requires simultaneous domination of the environment and suppression of the other teams' learning, and (3) how the use of team rewards can increase the probability that a team develops specialized agents that achieve these two goals simultaneously.

3. Environment, Agents, Teams and Cultures

Environment

We use the Gathering Game defined in Deep mind's 2017 paper on Sequential Social Dilemmas [3] as the environment for Team-based MARL. The game is originally structured as a partially observable Markov game for 2 agents. We improved it to allow teams of agents to play each other and to enable the agents to identify whether the other agents in their observation space are US (of the same team) or THEM (of different team).

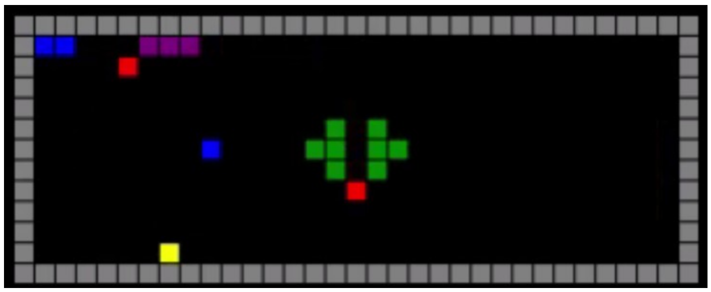


Figure 3: The Gathering Environment supports multiple teams of agents and provides them with an US versus THEM context.

Except for the improvements mentioned above, the environment is identical to the original Gathering:

- Multiple agents are organized into teams. For our experiments, we have 3 teams each with a distinct team color and a single random agent:
  - Team "Vikings" (agent 0, 1, 2) – blue
  - Team "Saxons" (agent 3, 4) – red
  - Team "Franks" (agent 5, 6, 7) – purple
  - Random agent (agent 8) – yellow
- An agent has 8 actions:
  - FORWARD = 0
  - RIGHT = 1
  - BACKWARD = 2
  - LEFT = 3
  - ROTATE_RIGHT = 4
  - ROTATE_LEFT = 5
  - LASER = 6
  - NOOP = 7
- The agent receives a reward of +1 when they eat a green apple (green pixel)
- The eaten apple regenerates itself after 15 game steps after it has been consumed
- By firing its laser, an agent can tag out all the agents (both US and THEM) that are in its observation space. The agent does not receive any reward for firing its laser or tagging out other agents.
- An agent does not receive any reward or penalty for being tagged, but is kept out of the game for 25 game steps, after which it is respawned.

Agents

The interaction between Gathering and the teams of agents is structured as a partially observable Markov game. The true state of the game from an agent's viewpoint is represented by 4 frames which identify:

1. Location of the apples
2. Location of the US agents (agents of the same team)
3. Location of the THEM agents (agents of different teams)
4. Location of the walls

The game's true state is only partially observable by each agent through an observation space of 10x20 pixels as shown in Figure 4 and 5.

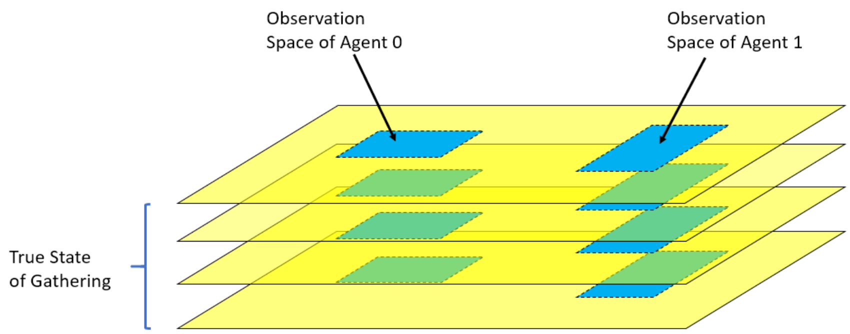


Figure 4: Each agent has an observation space which is a partial view of the true state of the Gathering environment.

Figure 5: An agent's observation space.

The agent intelligence is encoded within the agent's policy which is parametrized with a convolutional neural network (CNN) as shown in Figure 6. The 4x10x20 observation space of the agent returned by Gathering is inputted as a 4-channel input into the CNN. The 1^{st} convolutional layer take this and convolute it into an output of 16x10x20. The 2^{nd} convolutional layer then downsizes this output from the 1^{st} layer into an output of 16x5x10. The 3^{rd} and last convolutional layer further downsize the output from the 2^{nd} layer into an output of 16x3x8. This last output is stretched into a single vector of 384 and entered as input into a fully connected linear neural network with 8 outputs. These outputs are softmax-ed to arrive at the probability distribution for the 8 possible actions for the agent.

The agent's reinforcement learning is therefore based on the generic REINFORCE policy gradient algorithm which maximizes its individual reward. It does not require any policy parameters from other agents.

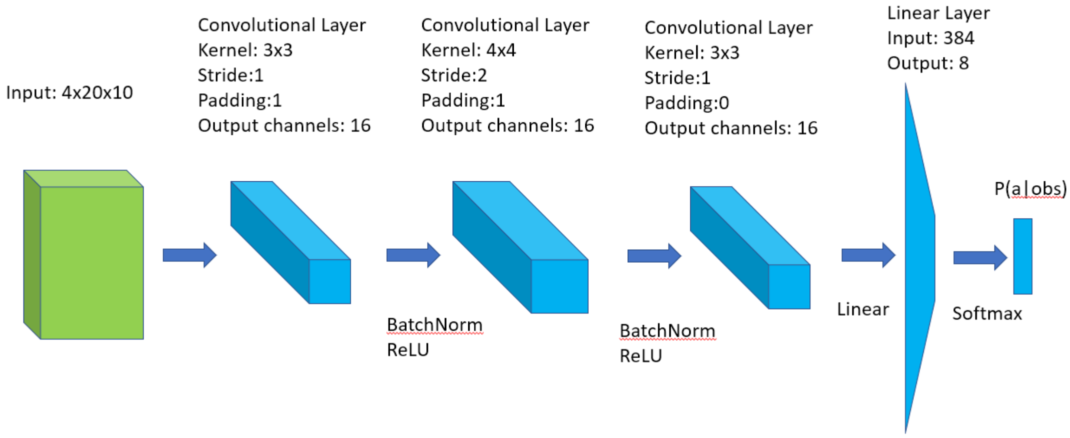


Figure 6: The agent's CNN policy network

Teams and Cultures

While it is very difficult to program intelligent agents to form organizations and develop organization intelligence, it is very simple to program the organization intelligence of the organization itself.

We encode a team's organization intelligence into its culture. In the reinforcement learning context, culture is how an organization doles out organizational rewards to its members on top of the rewards gathered by these agents from the environment, shown in Figure 7. Note that this team reward is only doled out during training, not during game plays.

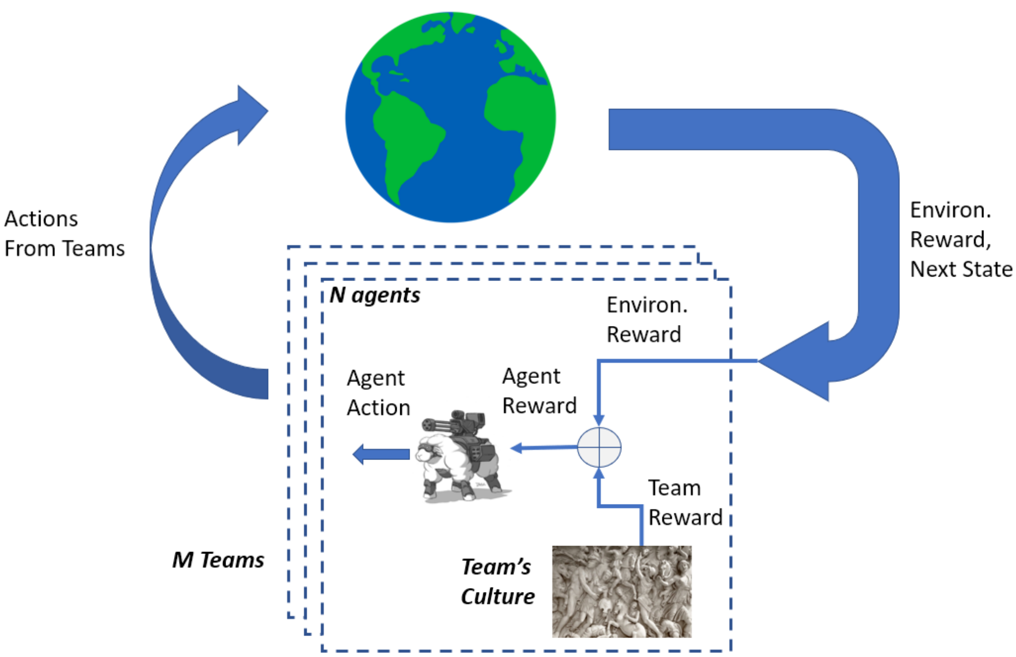


Figure 7: The Team doles out team rewards to its agents on top of their environmental rewards during training, but not during game play.

For the Gathering game, we have encoded 5 cultures for our teams of agents. These are described in Table 1:

Table 1: Team Cultures

Culture	**Team Reward Calculation**	**Parameter**
Individualist	$agent\ reward = environ\ reward$	None
Cooperative	$agent\ reward = environ\ reward + coop\ factor \times \frac{total\ team\ reward}{num\ team\ members}$	coop factor
No Fragging	$agent\ reward = environ\ reward - penalty \times friendly\ fire$	penalty per friendly fire incident
Warlike	$agent\ reward = environ\ reward - penalty \times friendly\ fire + reward \times enemies\ hit$	penalty per friendly fire incident; reward per enemy tagged out
Pacifist	$agent\ reward = environ\ reward - penalty \times laser\ fire$	penalty per firing of laser

The five cultures approximate the social behavioral concepts of individualism, collectivism, "no fratricide", militarism and pacifism. Note how these advanced concepts can be encoded into simple mathematical equations for team reward calculation in the RL context.

4. Experiments

Organization Intelligence

We conduct all our experiments with 3 teams of learning agents and 1 random agents jointly competing in the Gathering environment. The single random agent is colored yellow, while the 3 teams of learning agents have the following number of agents and colors:

- Team “Vikings” (agent 0, 1, 2) – blue
- Team “Saxons” (agent 3, 4) – red
- Team “Franks” (agent 5, 6, 7) – purple

For each of the 5 cultures (Individualist, Cooperative, no_Fragging, Pacifist and Warlike), we assign all teams (except the random agent) with that common culture, vary the cultural parameter and train them to either 2000 or 5000 game episodes. The cultural parameters and the number of episodes trained for each culture are presented in Table 2. For example, teams with Pacifist culture are trained to 2000 episodes with its cultural parameter “penalty for firing laser” set to -0.01, -0.1, -1.0, -10 and -100.

Table 2: Cultural Parameter Optimization and Training Episodes

	Cultural Parameters	Episodes Trained
Individualist	NA	5000
Pacifist	penalty=-100, -10, -1.0, -0.1, -0.01	2000
No_Fragging	penalty=-100, -10, -1.0, -0.1, -0.01	2000
Cooperative	coop_factor=0.01, 0.1, 1.0, 5.0, 10.0, 15.0, 20.0, 25.0, 50.0	5000
Warlike	penalty=1.0, reward=1.0	2000
	penalty=1.0, reward=0.5	2000
	penalty=1.0, reward=0.1	2000
	penalty=1.0, reward=0.05	5000
	penalty=1.0, reward=0.01	5000
	penalty=1.0, reward=0.005	5000
	penalty=1.0, reward=0.001	2000
	penalty=1.0, reward=0.0001	2000

We reason that if all 3 teams competing in Gathering have the same culture, the total rewards generated by these teams is the result of the combination of the agents’ agent intelligences and the teams’ organization intelligences specific to that culture.

We use the total reward generated by Individualist agents as the baseline for comparing between the different cultures. In Individualist culture, even though agents are assigned to teams, they act like lone wolfs because their agent rewards depend solely upon the environmental rewards they gather individually.

$$agent\ reward = environ\ reward$$

In Cooperative culture, the agent’s reward is the sum of the environment reward and the team reward doled out by its team.

$$agent\ reward = environ\ reward + \ coop\ factor \times \frac{total\ team\ reward}{num\ team\ members}$$

We therefore reason that the difference between the total rewards generated by Cooperative teams and that generated by Individualist agents is due to the team rewards doled out by the Cooperative teams. This difference can be used to quantify the organization intelligences of these Cooperative teams.

Are MARL Games Intransitive?

In single-agent reinforcement training (SARL), the end goal is for the agent's policy or Q network to learn the dominant strategy over the environment. In MARL, we have multiple agents or teams of agents simultaneously learning about the environment and about each other.

Thus, a very important question for MARL is whether the game in multi-agent setting has a dominant strategy or whether it is intransitive (like rock-paper-scissors). If the former is true, the end goal of MARL is still about getting the agents or teams of agents to learn the dominant strategy. If the latter is true, then MARL needs to get these agents and teams of agents to learn a portfolio of strategies and how one strategy can be used to overcome the other.

By reviewing the learning curves and the agent behaviors of the Individualist agents and those of the teams with Cooperative, Warlike, Pacifist and No-Fragging cultures, we are able to gain a very valuable insight into this important question.

Rewards or Penalties?

The 4 cultures - Pacifist, No_Fragging, Cooperative and Warlike - calculate their team rewards based on penalty, reward or a combination of both. These penalties and rewards are in turn based on a specific action (which the agent has full control), a 1st order result (for which the agent can achieve with a specific action) or a 2nd order result (for which the agent cannot directly achieve through its actions). This relationship between culture, team reward, reward vs penalty is summarized in Table 3.

Table 3: Penalty vs Reward in Cultures' Team Reward Calculation

Culture	Team Reward Calculation	Reward vs Penalty
Pacifist	$agent\ reward = environ\ reward - penalty \times laser\ fire$	Penalty for an action (firing the laser)
No Fragging	$agent\ reward = environ\ reward - penalty \times friendly\ fire$	Penalty for a 1st order result (tagging out US agent by firing the laser)
Warlike	$agent\ reward = environ\ reward - penalty \times friendly\ fire + reward \times enemies\ hit$	Penalty for a 1st order result (tagging out US agent by firing the laser) Reward for a 1st order result (tagging out THEM agent by firing the laser)
Cooperative	$agent\ reward = environ\ reward + coop\ factor \times \frac{total\ team\ reward}{num\ team\ members}$	Reward for a 2nd order result (total environ. rewards gathered by all members of the team)

We study the effects of rewards and penalties on agent behaviors by observing recorded videos and reviewing statistics of teams and agents (such as number of times the laser has been fired, number of US agents versus THEM agents tagged) gathered in repeated game plays. We will claim that the use of reward (as opposed to penalty) can result in agent specialization, which then enables a team to device a dominant strategy for a MARL game.

5. Results and Discussions

Organization Intelligence

In this section, we try to address the question "If every agent in the world was a [insert culture], how better would that world be compared to one where everyone is an Individualist?"

Table 4 summarizes the min and max of the total rewards generated by all the teams competing in Gathering based on culture (rows) and the number of episodes trained (column). The min and the max are computed from detailed training data in Table 7 of Supplementary Materials. In Table 5, we present the uplifts that teams with Pacifist, No_Fragging, Cooperative and Warlike culture have over the Individualist agents. These uplifts are quantitative measures of these teams' organization intelligences.

Table 4: Total rewards by all teams based on culture and episodes trained

	Ep = 1000	Ep = 2000	Ep = 3000	Ep = 4000	Ep = 5000
Individualist (baseline)	**271.5**	**280.6**	**344.2**	**355.1**	**375.4**
Pacifist – Max	537.5	570.0			
Pacifist – Min	479.8	511.0			
No_Fragging – Max	507.7	518.5			
No_Fragging – Min	390.7	396.0			
Cooperative – Max	394.9	427.7	573.5	613.2	614.0
Cooperative – Min	***262.6***	301.3	355.2	***338.5***	***353.3***
Warlike – Max	457.3	436.0	420.3	403.5	389.2
Warlike – Min	***0.0***	***0.0***	***0.0***	***0.0***	***0.0***

Table 5: Uplifts over Individualist by culture and episodes trained

	Ep = 1000		Ep = 2000		Ep = 3000		Ep = 4000		Ep = 5000	
Pacifist – Max	266.0	98%	289.4	103%						
Pacifist – Min	208.3	77%	230.4	82%						
No_Fragging – Max	236.2	87%	237.9	85%						
No_Fragging – Min	119.2	44%	115.4	41%						
Cooperative – Max	123.4	45%	147.1	52%	229.3	67%	258.1	73%	238.6	64%
Cooperative – Min	***-8.9***	***-3%***	20.7	7%	11.0	3%	***-16.6***	***-5%***	***-22.1***	***-6%***
Warlike – Max	185.8	68%	155.4	55%	76.1	22%	48.4	14%	13.8	4%
Warlike – Min	***-271.5***	***-100%***	***-280.6***	***-100%***	***-344.2***	***-100%***	***-355.1***	***-100%***	***-375.4***	***-100%***

It is not surprising that the Pacifist and No_Fragging cultures provide immediate and sizeable uplifts over the Individualist culture. By placing a penalty on firing laser or tagging a fellow team-mate, an agent learns very quickly that it can maximize its total reward by reducing or abstaining from firing its lasers. Since the agent no longer wastes a game step to fire its laser, it can focus more on gathering apples.

Furthermore, the reduction in aggressiveness generates a virtuous cycle. Over time, there is less and less need for agents of one team to tag out agents of the other teams, since they pose less and less of a threat. In short, peace is a good thing and the ability of all teams to reduce aggression by placing a penalty on either the action itself or on the action's 1st order result is an indication of organization intelligence. The uplifts on total rewards are immediate (after less than 1000 episodes of training) and universal (across a wide range of cultural parameters), as shown in Figure 8.

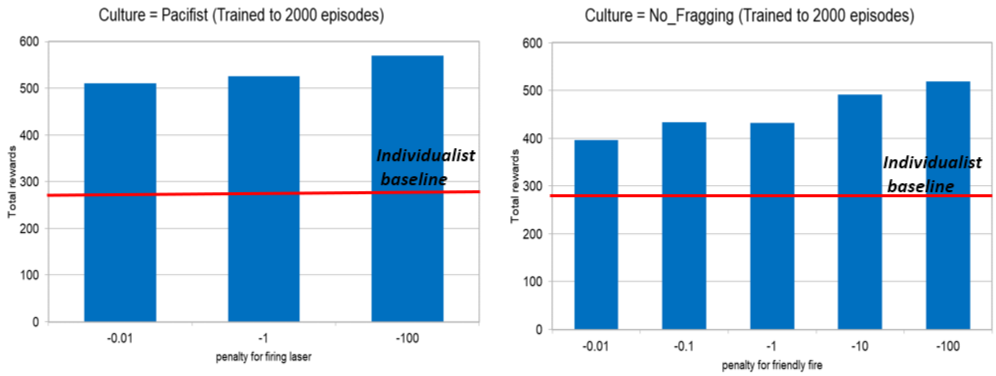


Figure 8: Average total rewards for Pacifist and No_Fragging cultures over a wide parameter range. Both are substantially higher than the Individualist baseline of 280.6.

On the other hand, the Warlike culture doles out team reward to an agent for tagging out an agent of a different tribe, which induces the teams of agents to immediately increase their aggressiveness very early on in the game (https://youtu.be/EQNNXn80iDQ).

$$agent\ reward = environ\ reward - penalty\ \times friendly\ fire + reward\ \times enemies\ hit$$

As shown in Figure 9, any Reward for enemy hits higher than +0.1 induces the teams to engage in mutually destructive firefights at the very positions they are spawned. In this continuous firefight lasting all 1000 game steps, none of the agents can even ventures beyond its spawn position to gather a single apple, resulting in total reward of zero. Only when the reward is reduced under +0.1 do the teams start behaving like teams with No_Fragging culture.

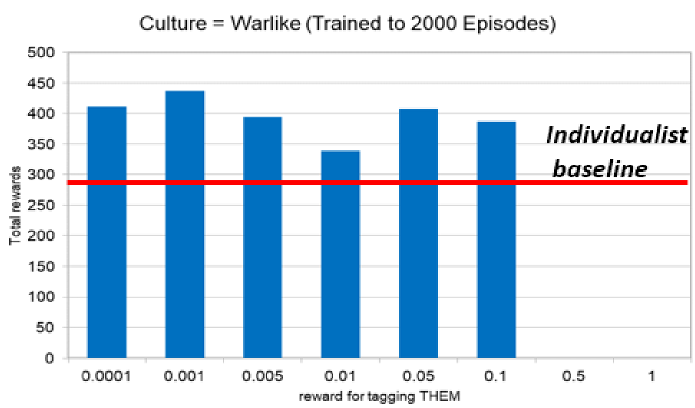


Figure 9: Average total rewards for Warlike cultures over a wide parameter range. Average total reward for Warlike teams is zero if the parameter "reward for enemy hit" is greater than 0.1.

The Cooperative culture doles out team reward to an agent based on a percentage of the total environmental rewards gathered by all the agents in its team. This percentage, the *coop_factor*, is its cultural parameter.

$$agent\ reward = environ\ reward +\ coop\ factor \times \frac{total\ team\ reward}{num\ team\ members}$$

The team reward for the Cooperative culture is similar to the company bonus an employee receives every year based on how well the company performs. It is thus based on a 2nd order result, meaning there is no clear and direct relationship between the desired result and the specific agent's action. We performed an exhaustive cultural parameter optimization search (*coop_factor*=0.01 to 50) to discover the optimal range for *coop_factor*, as shown in Figure 10.

We discover that using team reward based on 2nd order results can result in agent specialization and agent freeloading. In addition, we discover that using the right *coop_factor* (15.0 and 25.0) increases the probability that a team's agents specialize in a specific way which enables the team to dominate the game.

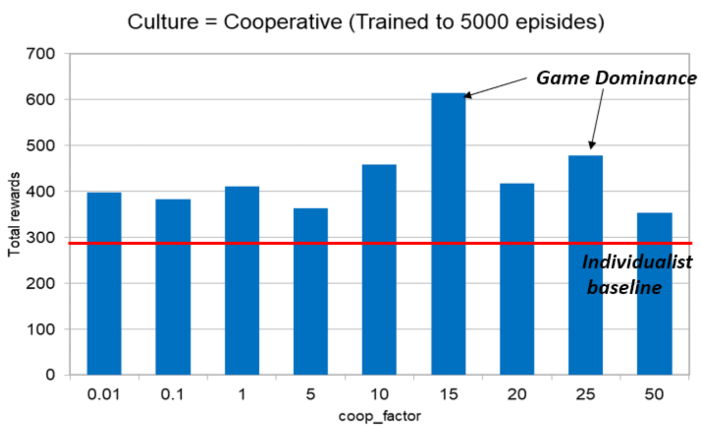


Figure 10: Average total rewards for Cooperative cultures over a wide parameter range.
Agent specialization leading to game dominance observed when *coop_factor* =15.0 and 25.0

Are MARL Games Intransitive?

In this section, we explore question of a different nature - "In a multi-agent world, how can one team gain a lasting edge over the other teams?"

The conventional approach to answer this question is a round-robin tournament where we pit teams of one culture against teams of other cultures in all possible permutations. However, this further raises the question of how these teams should be trained in the first place, and whether there should be teams of different cultures in the training, what the team sizes of these teams should be, so on and so forth.

Luckily, we were able to arrive at the answer through direct observations of agent behaviors and analysis of agent and team statistics. We venture to claim that even if a MARL game is intransitive to multiple individual agents, a team of agents with the right culture can develop specialized agents to completely dominate the game.

In SARL, many papers have demonstrated the ability of a single agent to learn a policy or Q-network that can dominate multiple games [5]. In MARL, the individual agent needs to learn a policy to dominate both the game environment as well as other learning agents. At a specific time, one of these agents may form a policy that dominates both, but over time the other learning agents will learn new policies to overcome this dominant policy. This is why it is hard for an individual agent to permanently hold its dominance in multi-agent games.

A team of agents however can develop specialized agents using team rewards. It can develop one set of agents to dominate the game environment, and a second set of agents to suppress the learning of the other teams' agents. When the other agents no longer learn, the team can permanently dominate the game.

For individual agents – Gathering is intransitive

When analyzing how Individualist agents learn, it became apparent to us that the Gathering game is intransitive. As shown in Figure 11, the top 3 scoring agents change continuously as training progresses. The individual agents are continuously adjusting and improving their policies against each other.

	Ep=1000	Ep=2000	Ep=3000	Ep=4000	Ep=5000
Agent 0	0	0	0	13	9.3
Agent 1	69.7	75.8	83.6	107.5	87.9
Agent 2	30.6	29.2	63.4	24.4	18.1
Agent 3	0	40.4	40.5	44.1	44.5
Agent 4	0	6	13.1	9.9	17.5
Agent 5	30.1	41	43.4	60.7	65.7
Agent 6	84.5	55.2	70.3	57	60
Agent 7	56.7	33	29.9	38.5	72.5

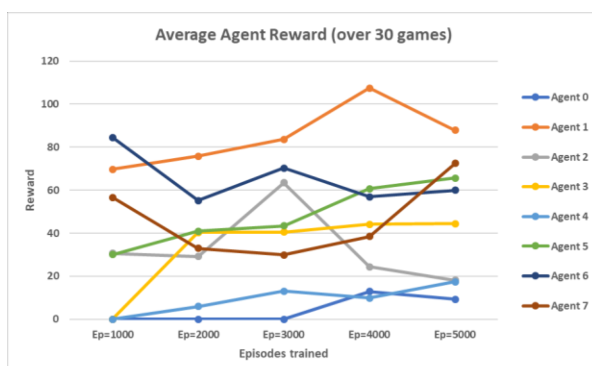
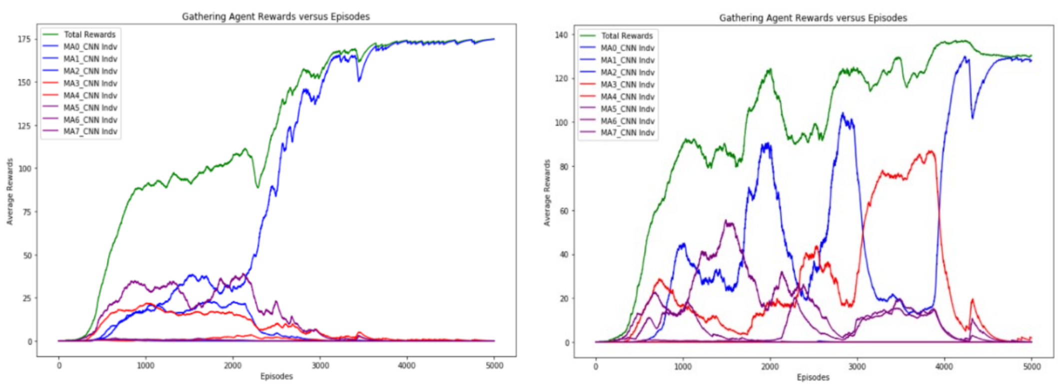


Figure 11: (Left) The top 3 score leadership, highlighted green, change hand between different agents throughout the training process. (Right) Naive agents continuously adjust and improve their policies based on the environment and on each other.

In the video showcasing these agents' evolving policies (https://youtu.be/MudcPMyog5Q), we can observe the Individualist agents continuously adapting and changing their tactics in order to overpower both the environment and one another. As long as the other agents are allowed to learn, one single agent cannot permanently dominate the Gathering game. Thus, the game is intransitive to the multiple individual agents.

Teams with specialized agents can dominate Gathering

The learning curves of Cooperative teams with coop_factor=15.0 and 25.0 are showcased in Figure 12. In both cases, Team Viking (blue) is able to dominate the game. The aggregate total rewards at Episode 5000 is entirely generated by Team Viking. All the other teams and their agents generate zero reward.



Coop_factor = 15.0

Average Statistics in Aggregate

Total rewards gathered	***614***
Num laser fired	999
Total US Hit (friendly fire)	0
Total THEM Hit	240
friendly fire (%)	0

Average Statistics by Tribe

Tribe Vikings has total reward of 614.0
Tribe Saxons has total reward of 0.0
Tribe Franks has total reward of 0.0

Coop_factor = 25.0

Average Statistics in Aggregate

Total rewards gathered	***477.1***
Num laser fired	1017.4
Total US Hit (friendly fire)	0
Total THEM Hit	245.8
friendly fire (%)	0

Average Statistics by Tribe

Tribe Vikings has total reward of 477.1
Tribe Saxons has total reward of 0.0
Tribe Franks has total reward of 0.0

Figure 12: Learning curves of agents with Cooperative culture when *coop_factor*=15.0 (top left), and *coop_factor*=25.0 (top left); aggregate and team statistics for *coop_factor* = 15.0 (bottom left) and *coop_factor* = 25.0 (bottom right). Team Viking has dominated the game by collecting all the rewards by episode 5000.

Team Viking permanently dominate the game because its agents have specialized. As shown in Figure 13 and 14, Agent 1 has specialized into an apple gatherer focusing only on maximizing the environmental reward, while Agent 2 has specialized into a warrior focusing only on tagging out agents of the other teams. By simultaneously dominating the environment (Agent 1) and suppressing the learning of the other teams (Agent 2), Team Viking has developed a permanent dominating strategy for Gathering.

Their technique can be observed in the following video (https://youtu.be/2u9SN0EoQYo). It is interesting to note that Agent 0 is the team's freeloader, walking around doing absolutely nothing and living off of the team reward. This is an unavoidable side-effect of the Cooperative culture.

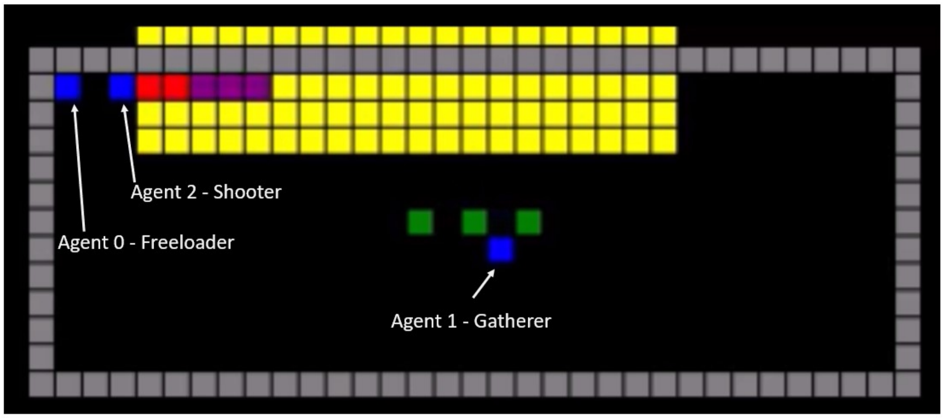


Figure 13: Agent Specialization and Freeloading in Team Viking (Cooperative Culture)

Figure 14: Agent statistics of Cooperative culture when *coop_factor*=15.0 (Left), and *coop_factor*=25.0 (Right). In both cases, Agent 1 collects all the apples while Agent 2 fires all the lasers in the game. The agents of the other teams (Agent 3-7) are completely suppressed - unable to either fire their lasers or collect any apple. Interestingly, Agent 0 becomes the freeloader of Team Viking, walking around doing nothing.

Rewards or Penalties?

In this section, we explore the question "How can we encode a team's culture so that it achieves our set goals?"

When we encode a team's culture, we have 2 sets of choices to make:

1. Do we use reward or penalty?
2. Should they be tied to actions, 1st order results or 2nd order results?

This is an important subject matter for future work, but here are some guidelines that we have induced from the experiments we have conducted so far.

Agent specialization requires reward

If the goal is for a team to build a dominating strategy to overpower both the environment and the other agents, it requires specialized agents. Agents can only be induced to specialize in performing tasks that do not generate environmental rewards if they are given team reward instead. The agent that specializes into a "shooter" in Team Viking does so because it maximizes on the team bonus:

$$coop\ factor \times \frac{total\ team\ reward}{num\ team\ members}$$

It learns that if none of the other teams' agents are allowed to escape out of its field of fire, the "gatherer" agent will gather more apples and it will in turn receive a higher team bonus.

Penalty stops harmful behaviors

Both the Pacifist and the No_Fragging cultures are based on penalties. The Pacifist's penalty based on the agent's action of firing its laser:

$$agent\ reward = environ\ reward - penalty\ \times laser\ fire$$

While the No_Fragging culture's penalty is based on the 1st order result of the agent firing its laser and in the process tagging out a fellow team-mate:

$$agent\ reward = environ\ reward - penalty\ \times friendly\ fire$$

The introduction of laser into the Gathering game renders it a prisoner's dilemma for all agents competing in the game [3], the aggressiveness of the agents elevates with increasing training because of Fear and Greed - the Fear of being tagged by the other agents and the Greed of being able to collect the apples by itself [6].

By placing a penalty on the use of the laser in the case of the Pacifist culture or on one of its unintended consequence (friendly fire) in the case of the No_Fragging culture, the agents learn to reduce or abstain completely from firing their lasers. This then leads to a virtuous cycle reversing the escalating aggression observed in the DeepMind paper.

Should Reward be based on Action or Result?

If we knew ahead of time what the team's dominating strategy is for a game, we can explicitly program the team's culture using reward and penalty based on specific actions.

As shown in Figure 13, Team Viking's dominating strategy requires only 2 specialized agents. The first agent, positioned on the left of all the other teams, needs to fire its laser throughout the game. We can give this agent a team reward based on the specific actions of orienting itself to the right and firing its laser the entire time. The second agent needs to go out and gathers apple. The Pacifist agents are very good at that, so we can give this agent a penalty based on not firing its laser and it will go out learning to gather apples.

In most cases however, we do not know ahead of time what the team's dominating strategy is. Setting a team reward based on 2^{nd} order results which equates to our set goal allows the teams of agents to search the team strategy space for us. As we have illustrated in the Cooperative culture example, through extensive cultural parameter optimization search we can create learning trajectories whereby some of these teams will develop the agent specialization necessary to arrive at the effective strategies for achieving our goals.

8. Conclusion and Future Works

In summary, we have demonstrated that we can vastly simplify MARL by separately encoding agent and organization intelligences, organizing agents under teams and providing them with US versus THEM context.

We also provide examples and reasonings why a multi-agent game can be intransitive to multiple individual agents not organized under teams, yet a team of agents can permanently dominate the game through agent specialization.

Furthermore, we have showcased how simple mathematical equations for calculating team rewards can be used to encode advanced social behavioral concepts into team culture, and how the use of penalties and rewards can shape agents into very effective teams capable of fulfilling a variety of set goals from maximizing the common goods to finding a dominating strategy to a game.

There are 2 immediate possible extensions to Team-based MARL. First is to apply it to a larger version of Gathering with many more teams and agents. Second is to apply it to a team-version of Wolfpack, where both the predators and the preys are organized under teams. These extensions can help us gain better insights into building more effective teams, achieving higher level of agent specialization and developing even more effective team strategies.

From an architectural standpoint, it may make sense to experiment with multi-cultural teams or hierarchical teams, and to compare their effectiveness to the single-culture teams that we have been using in this paper. It may also be very interesting to explore if it makes sense to parameterize the team culture itself using a linear or logistic regression model or even a deep network, so that we can automate the learning of the optimal team culture for a game or problem.

Team-based MARL also makes it very easy for researchers to explore how AI agents can enhance or protect human agents in a team-based setting. A game where multiple AI agents are organized around human players (e.g. protect the king or follow the leader) can be very useful for human-machine interaction researches. The human protection or enhancement roles of the AI agents can be programmed into the teams' cultures. During training, these AI agents can learn how to work effectively with human players. Later on, we can evaluate their true performance during game plays.

References

[1] Foerster, J. N., Chen, R. Y., Al-Shedivat, M., Whiteson, S. Abbeel, P., and Mordatch, I. Learning with opponent learning awareness. In AAMAS, 2018.

[2] Hu, J., & Wellman, M. P. (1998). Multiagent reinforcement learning: Theoretical framework and an algorithm. Proceedings of the Fifteenth International Conference on Machine Learning (pp. 242-250)

[3] J. Z. Leibo, V. F. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel. Multi-agent reinforcement learning in sequential social dilemmas. CoRR, abs/1702.03037, 2017.

[4] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. arXiv preprint arXiv:1706.02275 (2017).

[5] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, “Asynchronous methods for deep reinforcement learning,” CoRR,vol. abs/1602.01783, 2016.

[6] https://goo.gl/w2VqlQ

Supplementary Materials

3 teams of common culture and cultural parameter and 1 random agent compete in the Gathering environment with 1000 game steps per episode. Total rewards per episode is calculated by adding up the environmental reward of all the agents (note that there is no team reward during game play). Average total rewards are then calculated by averaging the total rewards per episode for 30 episodes of game play. The average total rewards by culture/parameter (row), and episodes trained are presented in Table 7.

Table 7: Average total rewards over 30 game plays by culture/parameter and episodes trained.

	Ep = 1000	Ep = 2000	Ep = 3000	Ep = 4000	Ep = 5000
Individualist (baseline)	271.5	280.6	344.2	355.1	375.4
Pacifist (penalty=-100)	537.5	570.0			
Pacifist (penalty=-1.0)	479.8	525.1			
Pacifist (penalty=-0.01)	510.8	511.0			
No_Fragging (penalty=-100)	507.7	518.5			
No_Fragging (penalty=-10)	499.7	491.5			
No_Fragging (penalty=-1.0)	505.7	431.8			
No_Fragging (penalty=-0.1)	390.7	432.7			
No_Fragging (penalty=-0.01)	402.1	396.0			
Warlike (p=-1.0, r=1.0)	0.0	0.0			
Warlike (p=-1.0, r=0.5)	0.0	0.0			
Warlike (p=-1.0, r=0.1)	363.7	386.6			
Warlike (p=-1.0, r=0.05)	446.7	408.2	335.1	403.5	382.7
Warlike (p=-1.0, r=0.01)	432.7	338.9	363.1	380.7	376.3
Warlike (p=-1.0, r=0.005)	369.6	393.7	420.3	384.6	389.2
Warlike (p=-1.0, r=0.001)	457.3	436.0			
Warlike (p=-1.0, r=0.0001)	400.2	411.3			
Cooperative(coop_factor=50)	361.4	387.8	380.8	391.1	353.3
Cooperative(coop_factor=25)	327.4	427.7	424.6	488.5	477.1
Cooperative(coop_factor=20)	333.8	427.7	457.2	449.8	416.2
Cooperative(coop_factor=15)	300.5	377.1	573.5	613.2	614.0
Cooperative(coop_factor=10)	272.7	373.8	443.7	383.9	457.4
Cooperative(coop_factor=5.0)	296.7	301.3	355.2	357.3	363.3
Cooperative(coop_factor=1.0)	262.6	386.2	413.7	408.9	410.5
Cooperative(coop_factor=0.1)	313.4	343.4	385.6	410.2	382.4
Cooperative(coop_factor=0.01)	394.9	334.8	378.0	338.5	397.0