Gifting in Multi-Agent Reinforcement Learning

Andrei Lupu
Mila, McGill University
Montreal, Canada
andrei.lupu@mail.mcgill.ca

Doina Precup
Mila, McGill University
Montreal, Canada
dprecup@cs.mcgill.ca

ABSTRACT

Multi-agent reinforcement learning has generally been studied under an assumption inherited from classical reinforcement learning: that the reward function is the exclusive property of the environment, and is only altered by external factors. In this work, we break free of this assumption and introduce peer rewarding, in which agents can deliberately influence each others' reward function. We formalize this more general setting and discuss its properties in depth. We also empirically study gifting, a peer rewarding mechanism which allows agents to reward other agents as part of their action space. We demonstrate that this approach can greatly improve learning progression in a resource appropriation setting and provide a preliminary analysis of the complex effects of gifting on the learning dynamics.

KEYWORDS

Learning agent-to-agent interactions, Multi-agent learning, Reinforcement learning, Reward structures for learning

ACM Reference Format:
Andrei Lupu and Doina Precup. 2020. Gifting in Multi-Agent Reinforcement Learning. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 9 pages.

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) extends standard reinforcement learning to the more realistic scenario where multiple learning agents must explore and exploit a shared environment [3]. MARL tasks can vary from fully competitive to fully cooperative, with many variations in terms of the set of allowable interactions between agents [16].

As a more general setting, multi-agent reinforcement learning violates one of the core assumption behind a majority of single-agent methods: due to agents learning and changing their behaviour throughout the training, MARL is intrinsically non-Markovian: the distribution of future states and rewards no longer depends only on the current state and action, but also on the entirety of the training process for each of the other agents in the environment. The Markov property could then only be maintained if the internal state of every agent were made public – an infeasible requirement in many cases, such as competitive settings. For that reason, MARL tasks often require optimizing a policy for an ever-changing setting, and come with their unique set of challenges, including possible feedback loops between agents that drive the system away from optimal collective behaviour [2].

In order to prevent further complications, previous multi-agent reinforcement learning studies have thus adhered to a simplification inherited from classical RL – treating rewards as coming exclusively from the environment. Although this function can be non-stationary [1] or even controlled to create a learning curriculum [15], it is assumed that its evolution from one training episode to the next is only due to an exterior process, beyond the control of the agents present within the environment itself. In other words, agents must learn to optimize the assigned reward function of their environment, but have no means of shaping it.

Whereas this assumption is natural in single-agent RL, it is an artificial restriction in multi-agent settings and results in agents learning to interact only insofar as it helps them maximize environmental returns. It also greatly restricts the type of interactions allowed between agents. For instance, it makes it impossible to simulate negotiations, gifts and trade deals, all of which are at the core of many interactions within our societies.

It is therefore natural to ask what dynamics arise when we relax the above assumption and make the reward function perceived by each agent no longer an exclusive attribute of the environment, but also dependent on *peer rewards*: rewards which originate from agents evolving in the same environment. As this introduces an even greater agent inter-dependence than merely having them learn in a common environment, we can reasonably expect many such models to provide a challenge to current MARL methods. However, just like non-stationarity can be used as a tool to design a learning curriculum, we argue that peer rewards can also be beneficial in accelerating the learning process and promoting stability in multi-agent systems.

To the best of our knowledge, this work is the first to investigate peer rewards in the multi-agent reinforcement learning setting. As such, we seek to lay the groundwork for future progress in this direction and thus define and analyse what is arguably the simplest peer rewarding mechanism: deliberate reward passing, or, in short, *gifting*.

The first motivation behind *peer rewards* is, as mentioned above, to free multi-agent reinforcement learning from the artificial restriction that rewards are exclusively sourced from the environment. In doing so, we allow for broader and more general settings of interest. Naturally, some of those settings will prove to be challenging but interesting tasks, which are necessary in paving the way towards a deeper understanding and broad application of reinforcement learning in the real world.

Beyond the challenge, we conjecture that peer rewards will also prove useful in training MARL systems to develop behaviours that would otherwise be difficult or nearly impossible to obtain.

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

The most obvious effect of enabling agents to reward each other is that the reward signal becomes more dense, which is known to generally facilitate learning [17]. In fact, this could lead to agents developing skills without any form of environment reward; something that is hinted at in the experiments which we will present later. Then, if agents successfully learn that the rewards they provide can influence other agents, they can use them to obtain a desired collective behaviour. In cooperative settings, peer rewards could for instance be used to encourage collaboration, or to decrease the cost of an agent self-sacrificing (i.e. incurring an individual penalty in favor of a collective gain). In competitive settings, peer rewards can be used to "buy" information from another agent or to strike a compromise. In either case, using peer rewards draws strong analogies to behaviour shaping, as introduced by Skinner in the field of experimental psychology, where an animal can be drawn to behave unnaturally by incrementally rewarding behaviours close to a target one.

Note that we also expect peer rewards to help with the exploration problem that is the crux of reinforcement learning. For instance, assuming agents can observe each other, an agent could reward others for visiting a state of which it is particularly uncertain, or that it suspects may be unsafe.

Finally, we hope that studying peer rewards in reinforcement learning will eventually help modelling and understanding social animals, in particular those with swarming behaviours, such as bees or ants.

2 BACKGROUND AND RELATED WORK

There exists a vast body of research on multi-agent reinforcement learning, with each work addressing one of the many challenges of the field [16].

One such challenge tackled by MARL studies is how to enable agents to learn to coordinate to solve a task. This has often been achieved by having all agents maximize a common shared reward [6, 12, 13]. However, this method is limited to the fully cooperative setting, where agents have no opposing interests. As a result a lot of effort in recent years has been dedicated towards proposing more complex coordination methods, often by enabling agents with additional capabilities. For instance, in emergent communication studies, agents are allowed to share information, and must learn a common language to do so effectively [7]. In opponent modelling, agents are no longer treated as simple parts of the environment, but recognized as capable of learning and agency [8], which enables reasoning about other agents' behaviour and therefore can lead to more complex interactions.

Our paper utilises the Harvest environment introduced by Perolat et al., and we refer to their work throughout the paper since many of our findings naturally relate to theirs. Others have also sought to encourage coordination in the Harvest environment, but have done so through providing agents with intrinsic motivation based on their social influence, rather than a direct reward transfer mechanism between agents [11].

Part of our approach and discussion can be related back to curriculum learning, where an RL agent is training through a sequence of tasks that are selected to be progressively harder, with the goal of ultimately maximizing reward in a complex target task [15]. Although curriculum learning has typically been reserved to classical RL settings, it has also been successfully used for multi-agent ones [9].

Finally, we also draw parallels with works from the evolutionary biology field, in particular those concerned with reciprocal altruism between organisms [21, 22].

3 OUR APPROACH

As this is the first work to study peer rewarding, we strive for simplicity of design throughout our approach.

First, we avoid any form of agent modelling or inter-agent communication and enable agents to perceive each other only through observations from the environment. This allows us to isolate the effects that peer rewards have on the learning progression of the system, without interference from other methods that can alter behaviour (for instance, by actively promoting cooperation).

Second, simplicity was also behind the choice of peer rewarding mechanism to analyse in this work. We focus exclusively on *deliberate reward passing*, or *gifting* in short. We define a gift to be a reward given to an agent a_i as a direct and immediate effect of a deliberate action of an another agent a_j. In particular, the rewards that agent a_i obtains from receiving a gift are independent from those it receives by acting upon the environment. Note that, like any other action, gifting is mediated by the environment and so may be subject to limitations. For instance, agents may be able to gift only to agents within a given range or gift values might be restricted.

Finally, we are interested in emergent behaviour and in evaluating whether standard reinforcement learning methods can naturally learn to employ reward gifting. As a result, we do not prescribe when an agent should gift and instead elect to simply modify the capabilities of the agents by extending their action space $\mathcal{A}$ with a reward gifting action. The onus is therefore on the RL algorithm at hand to correctly explore and learn to utilize gifting.

3.1 Gifting Mechanics

Naturally, gifting must be actuated by the environment. Therefore, designing any gifting mechanic (or, more broadly, any peer rewarding mechanic) necessarily makes assumptions about the structure of the environment. For instance, if a gift is to be distributed to all agents within range, as is the case for the mechanics that are of concern here, it is implied that the environment has a notion of agent positioning.

Now we detail the three gifting mechanics which are analyzed in this paper.

Zero-Sum. This mechanic imposes no restriction on the number of times an agent can pick the gifting action within an episode. However, each time it elects to send a gift of value g, it incurs an immediate penalty of $-g$.

This type of gifting incorporates the principle that "what I give is what I lose", which is the essence of most of our daily monetary transaction. It also corresponds to reciprocal altruism observed in the animal kingdom, defined as a "*behaviour that benefits another organism [...] while being apparently detrimental to the organism* *performing the behaviour*", with the expectation of reciprocity from the receiving organism at a later time [22].

This mechanic is of particular interest because it assumes that both the gifting and the receiving agent value the gift equally, which makes the the transaction zero-sum from a game theoretic perspective. This also implies that no net reward is added to the total reward pool, and so the average obtainable return per agent remains unchanged regardless of the number of gifts being exchanged. Finally, the constant penalty means that the action is discouraged unless it indirectly provides an advantage.

Fixed Budget. In this case, agents are allocated a fixed "budget" of size B at the start of each episode, which is decremented by g each time the agent chooses to gift. Once the budget is depleted, gifting is no longer available as an action until the end of the episode. Note that in this case, agents cannot themselves draw rewards from their own gift budget and they are not penalized for gifting.

This draws direct inspiration from economic incentives that a government might give a company to encourage a specific corporate behaviour (such as to perform the switch to green technology).

In this case, the action of gifting increases the total reward pool. Thus, if agents select it with sufficient frequency and if B is large enough, the rewards obtained through gifting could potentially shift the focus of the system away from optimizing environmental rewards.

Replenishable Budget. This mechanic is similar to the above, but relies on the added assumption that environment rewards can be obtained throughout the episode, and not only at the end (upon reaching a goal). Each agent now starts with an initial budget $B = 0$, which is incremented as a function of the rewards it can collect from the environment. In this paper, we increment the budget by a fixed fraction of the environment rewards collected, albeit a stochastic scheme can also be easily implemented. Note that to prevent undesired feedback loops, we only increment an agent's budget if it collects a reward from the environment, not if it receives a gift.

Gifting with this mechanic also increases the total reward pool and so many intuitions are the same as for the fixed budget. However, it adds a dependence between rewards that agents collect from the environment and their capacity of gifting, which can potentially model scenarios where individuals become more influential as their fitness or possessions increase.

4 EXPERIMENTS

4.1 Setting

As a setting for the empirical study of deliberate reward passing, we choose the problem of common pool resource appropriation (CPR), where agents exploit a shared resource to their individual benefit, while being careful not to over-exploit and deplete the pool, thus hurting everyone's return. This problem was first studied in the reinforcement learning community through the *Harvest* environment introduced in Perolat et al. [18]. We re-implement this environment for the purpose of the paper, building on top of the *MiniGrid* repository [5].

In Harvest, seen in Fig. 1, there are 10 agents competing for collecting apples, each of which gives an immediate reward of 1 when eaten. Apples do regenerate, but at a rate dependent upon the number of remaining apples in the immediate vicinity. This means that if a region becomes entirely depleted at any given point, no more apples will grow there until the end of the episode. Therefore, in order to maximize their return, agents must learn to collect apples sustainably, without depleting the environment altogether. In addition, agents are equipped with a "laser beam", with which they can tag other agents. Once tagged, an agent is removed of the environment for 25 frames before being brought back in. Tagging another agent provides no reward to the agent that fired the beam, but momentarily reduces the competition it faces for collecting apples. Similarly, an agent receives no negative reward for being tagged, but cannot gather apples or receive any gifts until it returns to the game.

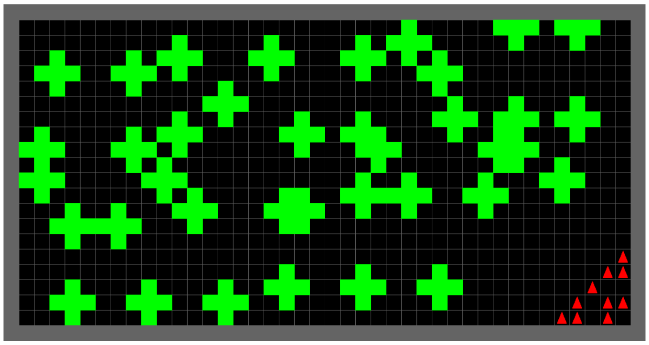

Figure 1: Our implementation of the Harvest environment. Apples are in green, agents in red. Agents employ embodied vision, meaning they only observe a limited are in front of them at any given time.

The agents in the work of Perolat et al. do ultimately converge to a collectively sustainable behaviour by tagging each other in order to reduce the effective population to the carrying capacity of the environment. However, they only do so after traversing a period of greedy behaviour for all agents, which is marked by a quick depletion of the resource pool and a very small return per agent. This period is called the "*tragedy of the commons*", and is a well known problem in economy [10].

Because artificial agents traversing a tragedy of the commons could have potentially disastrous consequences if deployed in a real resource exploitation scenario, and because CPR as a task presents a non-trivial mix of competition and cooperation, we consider the Harvest environment to constitute an excellent test case for the benefit of reward gifting in multi-agent reinforcement learning.

In addition to the laser beam, we also equip agents with a "gifting beam". Then, when an agent chooses the gift action, a gift of value $g = 1$ is equally split among all other agents within the beam range. Both beams have the same range, and the gifting is regulated according to one of the three mechanics described in section 3.1. Note that the gifting beam does not tag agents. Thus, if the laser beam consists of an indirect punitive action, the gifting beam is a direct rewarding action. For agents with a fixed budget, we set $B = 40$ and for those with a replenishable budget, we increased the budget by 1 for every 2 apples collected.

Agents perceive the environment through a first-person partial observation in the RGB domain (see [18] for details). For agents with a budget, their current budget is also provided through the observation array. Additionally, we choose to let actions be private and agents be indistinguishable of each other. This is done for scalability reasons, as keeping a different firing and gifting policy for each different agent becomes intractable.

Finally, for both simplicity and scalability to systems with a large number of agents, we steer away from agent modelling and let each agent be controlled by a Deep Q Network [14] – one of the simplest and best understood RL algorithms, and also the one previously used on Harvest by Perolat et al..

We intend on releasing the full code-base, including trained models, for open access in the upcoming months.

4.2 Evaluation Metrics

We use the same social metrics[1] to evaluate our agents as those used in Perolat et al. [18]. Let $\tilde{r}_i^t$ and g_i^t be the environmental rewards and the peer rewards obtained by agent i on step t respectively, and let $r_i^t := \tilde{r}_i^t + g_i^t$. Define $R_i := \sum_{t=1}^{T} r_i^t$ to be the return of the agent over an episode, where T is the horizon, set to 1000 in our experiments. Finally, let N be the number of agents in the environment ($N = 10$ in our case).

The four metrics are: the *average return (R)* per agent, *Sustainability (S)*, measuring the average time at which environmental rewards are collected, *Peace (P)* measuring the average number of agents that are untagged at any given point, and *Equality (E)*, given by 1 minus the Gini inequality index. Formally[2]:

$$R = \frac{1}{N}\sum_{i=1}^{N} R_i, \quad S = \frac{1}{N}\sum_{i=1}^{N} t_i, \text{ where } t_i := \mathbb{E}[t \,|\, \tilde{r}_i^t > 0],$$

$$E = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} |R_i - R_j|}{2N\sum_{i=1}^{N} R_i},$$

$$P = N - \frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{N} \mathbb{I}\{\text{agent } i \text{ timed-out on step t}\}.$$

4.3 Results

We study the behaviour of agents equipped with one of the three gifting mechanics described in section 3.1. We also reproduced the behaviour of agents devoid of gifting, matching closely the results reported in Perolat et al. [18], to serve as a comparison baseline. All curves are averaged over 4 sets of hyper-parameters, each evaluated over 10 episodes with randomized agent starting positions. The shaded area around each curve represents the maximum and minimum values observed.

Figures 2 and 3 respectively show the average return and sustainability obtained by each model. Both metrics are highly correlated, which is to be expected. All models have a very similar performance in the early stages of training (<1000 episodes). Past that mark, sustainability and return drop, with both the baseline and the two budgeted gifting models falling into a tragedy of the commons of equal magnitude and duration, lasting roughly between episodes 1500 and 3500. Agents equipped with zero-sum gifting also see a small performance decrease, most visible on Fig. 3, but recover almost immediately, avoiding the tragedy and converging to their optimal performance around the 2500 episode mark, nearly 2000 episodes earlier than the other models.

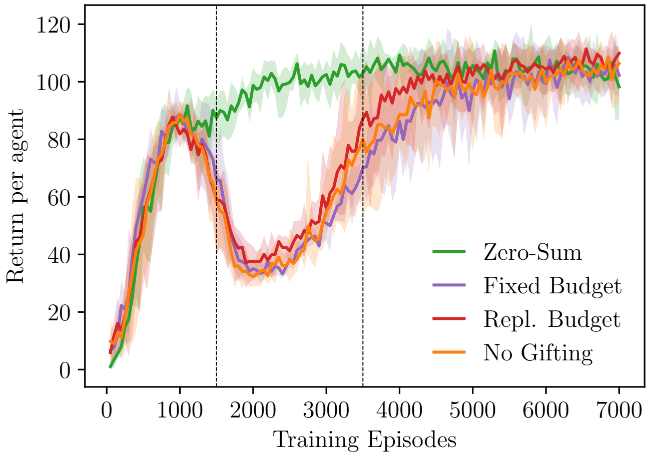

Figure 2: Average return per agent for the three gifting mechanics and baseline, plotted throughout training. Agents employing zero-sum gifting are the only ones suffering no return drop from the tragedy of the commons.

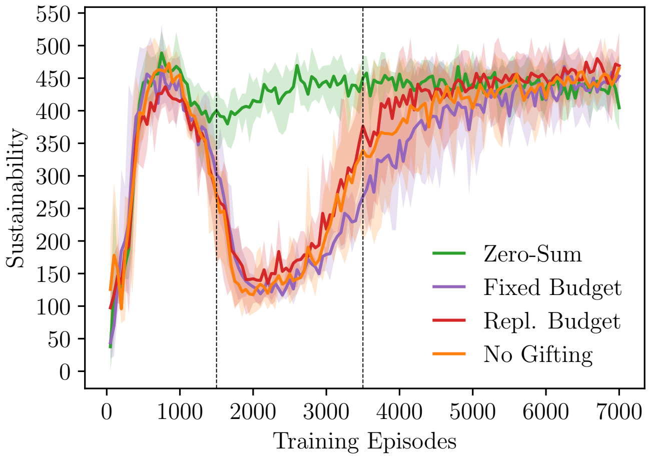

Figure 3: Sustainability for the three gifting mechanics and baseline, plotted throughout training. The zero-sum models see a small dip in performance before quickly recovering and converging to the final performance. The baseline and budgeted models only do so after suffering a great tragedy.

Perolat et al. found that agents devoid of gifting solve the tragedy of the commons and increase sustainability by resorting to using their tagging beam more often to reduce competition over apples. The system then stabilizes once the average number of untagged agents (the *Peace*) is brought down to the carrying capacity of the environment. We observe the same effect in in Fig. 4 of our results.

[1] We replace Efficiency by the average return, but both are equivalent up to a constant factor.

[2] $\mathbb{I}$ represents the indicator function.

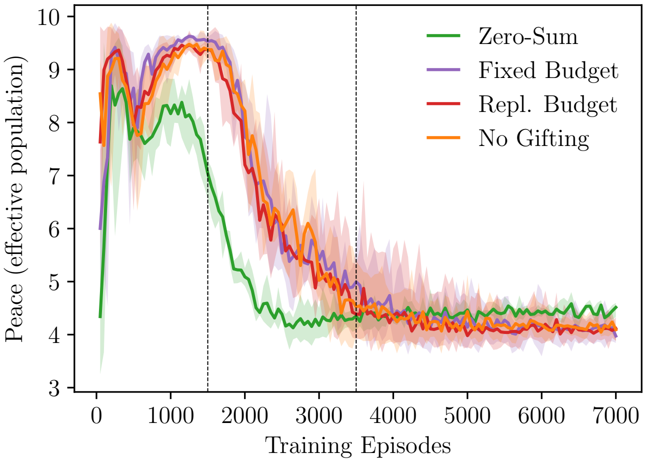

Figure 4: Peace metric for the three gifting mechanics and baseline, plotted throughout training. Zero-sum agents make active use of the firing beam much earlier in training, but all models converge to the same final value.

The curves for the budgeted models match the baseline closely, but the zero-sum exhibits greater conflict earlier in training, reaching a plateau much faster – around the same time as the agents converge to obtain their peak return. Also note that all models converge to the same peace value, indicating that the different gifting mechanics have little effect on the carrying capacity of the environment.

Finally, equality (Fig. 5) can be seen to be mostly unaffected by reward gifting, whether it is governed by a budgeted or zero-sum mechanic. The very low starting equality of agents with zero-sum gifting is only a transient effect resulting from agents sharing more gifts than the number of apples they collect. Once agents learn to be better gatherers and the gifting rate drops, the effect is quickly eliminated, as seen on the bottom plot of Fig. 5. Then, all systems reach the same level of return equality between the agents. Note that we do not expect perfect equality, since there can be a significant advantage granted by an agents' starting position.

5 DISCUSSION

The results in the previous section demonstrate the complex effects that peer rewards can have on the agents' learning progression and speed of convergence to an optimal behaviour. The experiments also demonstrate that careful consideration must be given to the design of the rewarding mechanisms, as they can lead to drastically different outcomes. In this section, we analyse the empirical results obtained, and dissect the impact of gifting on the learned behaviour.

It is obvious from section 4.3 that only agents equipped with the zero-sum gifting mechanism successfully avoided the tragedy of the commons. Indeed, both the fixed budget and replenishable budget gifting models were consistently subject to the same drop in performance observed in the baseline.

We are able to distinguish at least two ways in which zero-sum gifting differs from the other models, allowing the agents to avoid the tragedy of the commons: an early appreciation of tagging and a slowdown of the descent into a greedy behaviour.

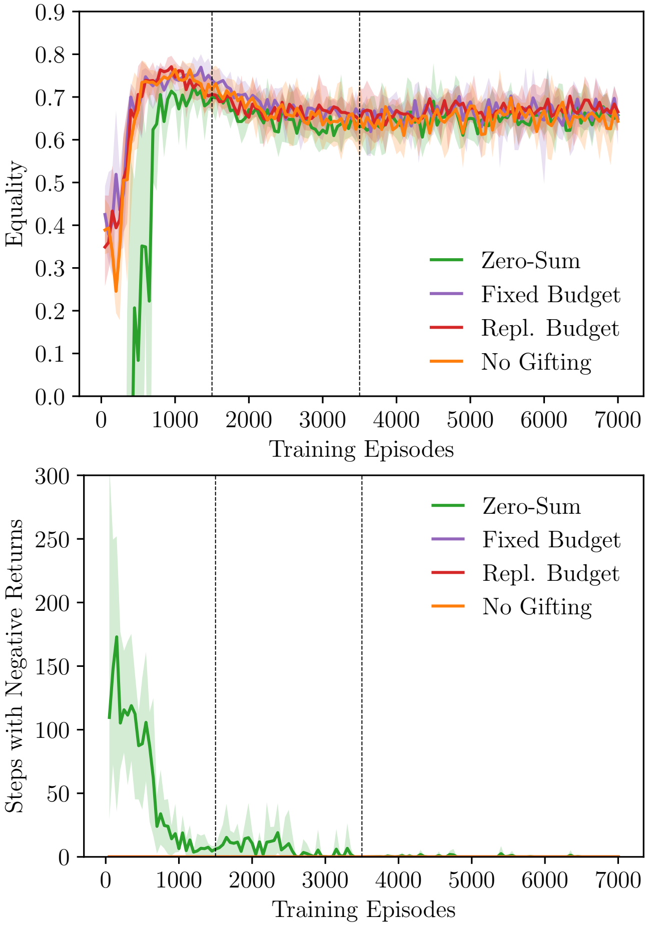

Figure 5: Top: Equality for the three gifting mechanics and baseline, plotted throughout training. Bottom: Average number of steps spend by zero-sum agents with a negative cumulative return up to that step. The low equality of the zero-sum model in early stages is due to agents gifting more than they gather and observing the resulting negative return. This is just a transient effect, and quickly all models converge to the same equality score.

Improved Tagging. As mentioned previously, systems of zero-sum agents see their Peace metric – an inverse measure of conflict – decrease significantly earlier than all other models. We believe this is due to an interplay between firing and gifting, and to the ε-greedy training method. Indeed, since we randomize actions during training, each agent will effectively be forced to gift at times, especially during the early training episodes (we start with $\varepsilon = 1$ and decrease it linearly). If there are other agents present within range, a reward is transferred, and the gifting agent incurs an immediate negative reward. Thus, because the tagging beam and the gifting beam share the same range, zero-sum agents may learn much earlier to tag each other, with the goal of removing agents in front of them and ultimately avoiding the penalty of a forced gift action. In other words, tagging for zero-sum agents is not only incentivized

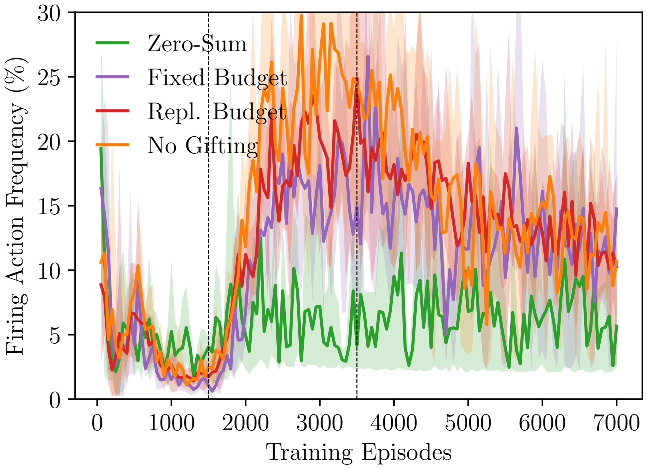

Figure 6: Firing rate for the three gifting mechanics and baseline. All models increase their firing rate when they begin encountering the tragedy of the commons. Firing rates then decrease as agents become more accurate with the beam.

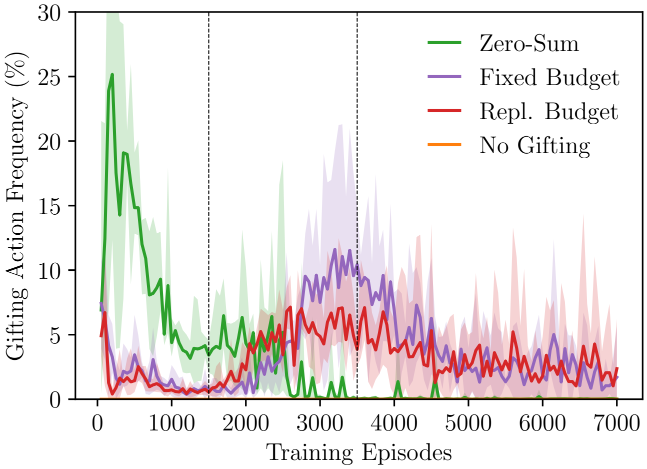

Figure 7: Gifting rate for the three gifting mechanics and baseline. All gifting rates ultimately approach zero as an effect of best-response dynamics.

as a way of reducing competition for apples, but also of avoiding negative rewards induced by forced gifting during training.

However, zero-sum gifting does not in fact render the agents more aggressive towards each other, despite what the Peace metric could indicate. Looking at Fig. 6 reveals that agents enabled with the zero-sum mechanic generally have a much lower firing rate than the other models. In fact, for a majority of the time corresponding to the tragedy period, the frequency at which zero-sum agents select the tagging action on average is approximately a quarter that of the baseline, and less than a half that of the two budgeted gifting models. Reconciliation between the firing frequency and the much lower Peace score of zero-sum agents can only be obtained if the latter are effectively more accurate when aiming. Thus, the zero-sum gifting mechanic enables agents to quickly learn to master the tagging beam by further incentivizing its use during the early stages of training. Consequently, agents become skilled with the tagging beam before it even becomes necessary to regulate the exploitation of the resource pool. This is therefore solid evidence that peer rewards can greatly allow for improved skill acquisition, possibly without even the need for environmental rewards.

An early mastery of the tagging beam is then one of the key phenomena that allow zero-sum agents to avoid the tragedy of the commons. Indeed, with this skill already acquired, zero-sum agents merely have to learn to use it more often once the return begins to drop as a result of greedier gathering. That knowledge is easily acquired, judging by the fact that, in Fig. 6, all models see a rise in their firing rate almost immediately after a decrease in returns. By contrast, the budgeted models and the baseline not only have to learn that tagging is necessary, but also *how* the tagging beam works (i.e. what its range is). That is the reason why those models see their firing rate spiking much higher after hitting the tragedy of the commons. Then, as learning progresses and returns stabilize, they too learn to use the tagging beam accurately, leading to a decrease in their firing rate, all while maintaining the effective population at the carrying capacity.

Note however that models with budgeted gifting still require a somewhat lower firing rate than the baseline to obtain the same effective population. This is in spite of them not benefiting from the same interplay between tagging and the gifting penalty, as described for the zero-sum mechanic. We attribute that to the fact that the tagging beam and the gifting beam share the same range. Indeed, we conjecture that the range of both beams is learned by the same part of the neural network controlling each agent, rather than attempting to learn each range separately. Then, whenever an agent elects the gifting action, it also simultaneously learns about the tagging range, regardless of if the gifting was successful. This provides a second piece of evidence in favor of considering peer rewards as a tool to accelerate skill acquisition.

Reduced Greed. Given the complex relationship between firing and gifting, it would be easy to rule it as the only explanations behind the usefulness of zero-sum gifting in avoiding the tragedy of the commons. However, having a closer look at Fig. 3 and 4 to simultaneously compare the sustainability and peace scores reveals a richer picture. Indeed, observe that, for an equal peace level, the return and sustainability of zero-sum systems can differ greatly from those of the baseline or the budgeted systems. This can be taken as a hint that zero-sum agents are not merely better at tagging, but also better at gathering sustainably, at least during the period of training where the Peace gap exists.

We can gain additional insight by evaluating the very agents studied in section 4.3 in a setting where both tagging and gifting are disabled. In doing so, we essentially remove all direct agent interactions to disentangle the agents' attitude regarding their own exploitation of resources from their tendency to regulate others. Because we did not build our networks to handle a variable number of actions, we simply take the argmax over the output layer, excluding entries corresponding to the tagging and gifting actions. Note that we do not retrain the models. We simply saved the models

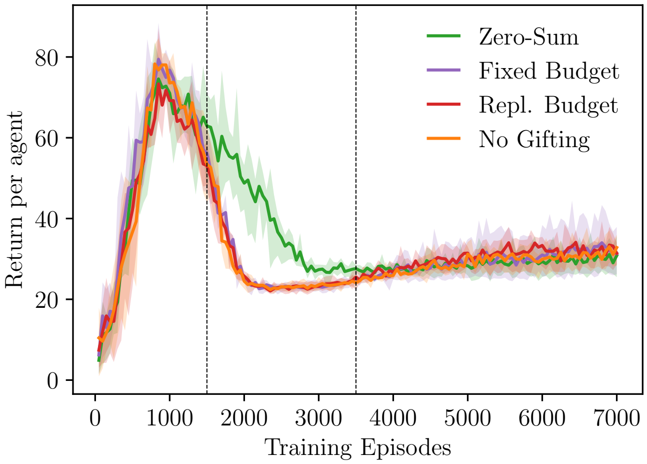

Figure 8: Return per agent when gifting and firing are disabled.

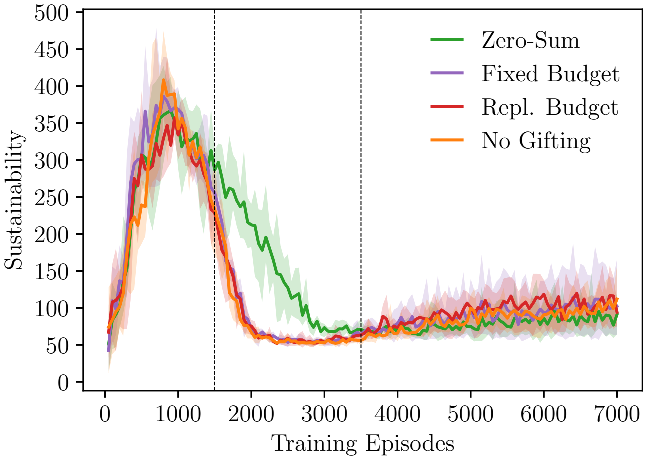

Figure 9: Sustainability when gifting and firing are disabled.

at regular interval during the initial experiments and re-evaluate them here.

As we can see in figures 8 and 9, all agents suffer greatly from not having access to the tagging and gifting beams, failing to recover a sustainable use of the resource pool after adopting a greedy behaviour. This is to be expected, since all systems overcome the tragedy of the commons in great part by learning to reduce the effective population to the carrying capacity of the environment. From a game theoretic perspective, the outcome observed in the no-interaction regime corresponds to the Nash equilibrium of the setting. Indeed, if any one agent was to reduce its own gathering to avoid depleting the resource pool, it would only be taken advantage of by the other agents, and its return would be ultimately lower. Thus, an aggressive exploitation of resources is a best response when tagging is not an option.

However, it can again be seen that the zero-sum gifting mechanic results in a different behaviour progression, with the descent towards greedy gathering being slowed down. Indeed, between episodes 1500 and 3000 of training, zero-sum agents have a more sustainable behaviour, effectively refraining at times from collecting isolated apples. Cross-checking with Fig. 7, we see that gifting for the zero-sum models reaches a plateau that coincides with its descent into greediness. We believe therefore that this slower descent observed in the no-interaction regime is attributable to agents using gifting during training to encourage restraint. As a result, agents learn how to better avoid depleting the resource pool, even though they ultimately do become greedy gatherers by following best-response dynamics.

A natural question to ask is: if zero-sum agents can slow down their descent into greediness, why couldn't agents with a budgeted gifting mechanic do the same? We believe the answer lies in how easy it is to learn the range and mechanics of each type of gifting, which is tightly linked to the feedback density for choosing the gifting action. In the case of the zero-sum mechanism, the gifting agent receives a negative reward if and only if there are others in range, allowing to quickly learn the workings of the beam. For budgeted gifting, there are no rewards (positive or negative) associated with gifting. Instead, the budget is simply incremented or decremented based on the actions taken. Additionally, if the budget is depleted, a gifting action is unsuccessful even if there are agents in range, further blurring the information received and making the range even harder to learn.

The fact that agents with a budget-based mechanism receive no reward whatsoever for gifting whereas those with zero-sum gifting do also means that they experience a different reward density. Since denser rewards – positive or negative – are known to facilitate reinforcement learning, this could also server as an additional explanation as to why zero-sum agents perform generally better and converge to their end behaviour much faster.

5.1 Links to Evolutionary Biology

We conclude this section by a discussion on the links between direct reward passing and the reciprocal altruism observed in the animal kingdom. Reciprocal altruism, as introduced by Trivers, encompasses any action by which an organism decreases its own immediate fitness to increase the fitness of another, with the expectation of reciprocity in the future [22]. The fitness is understood here in broad terms, but for the purpose of reinforcement learning, can be thought of in terms of rewards. Thus, an agent can be deemed altruistic if it voluntarily incurs a loss (or the risk of a loss) to allow another agent to observe a gain. By that definition, agents gifting with the zero-sum mechanism (and to a lesser extent with the budgeted mechanisms) exhibit an altruistic behaviour.

Looking at Fig. 7, it is obvious that all models ultimately refrain from gifting. Note that this holds for budgeted models as well, as although their respective curves do not go to zero, it is only because they occasionally use the gift action as a placeholder for waiting (no-op) when their budget is depleted. Intuitions from game theory can easily justify why agents stop gifting: it is a best response for any one agent to instead selfishly take advantage of others' generosity, all while focusing on gathering more apples. However, this raises the following question: what would it take for agents to not defect and instead converge to a common altruistic behaviour?

The complete answer likely depends on the dynamics of each environment, but Stephens [21] provides two key conditions that are necessary for reciprocal altruism to emerge. Those conditions were formulated in the context of the evolution of biological organisms, but translate naturally to reinforcement learning agents. The first condition is that a very large number of opportunities for altruism must exist, such as to have reasonable expectations of reciprocity. This is of course easily satisfied by the number of frames used in a typical RL experiment. The second condition, much less trivial, is that there must exist a mechanism by which to identify "cheaters" – agents reaping the benefits of others' altruism without reciprocating. This condition is necessary so that agents can learn to avoid being exploited by such cheaters, or even punish them to discourage selfish behaviours with regards to the peer rewards shared.

In the Harvest environment, this could for instance translate into each agent having an identifier that would distinguish it from all others. Such an approach would not scale, however, because it would then require to learn a different gifting policy for each distinct agent lying in the range of the gifting beam. The design of scalable mechanisms for detecting cheaters is therefore worth studying in future works.

6 CONCLUSION

In conclusion, this is the first work that studies the effects of peer rewards in multi-agent reinforcement learning. Our empirical results show that multi-agent systems can undergo multiple phases throughout their learning progression, corroborating the findings of Perolat et al.. We also find that the behaviour can vary drastically from one model to the next and that the characteristics of each phase, such as if the agents traverse a performance drop or not, is dependent upon the tools available and the level of mastery for each of them.

In particular, we demonstrate that deliberate reward passing – or gifting – can be a simple yet powerful way for both altering the learning progression of multi-agent systems and for helping them acquire skills that are a priori unrelated to the gifts. In particular, the zero-sum gifting mechanism, where an agent incurs a negative reward to distribute a positive reward of equal magnitude to nearby agents, proved to be the most influential in driving the evolution of the system's behaviour. Indeed, in the experiments performed on the common pool resource appropriation environment called Harvest, zero-sum gifting acted as both a facilitator for the acquisition of the tagging skill and as a way for agents to incentivize each other to maintain a sustainable gathering. Consequently, the agents successfully avoided the tragedy of the commons that plagued all the other models tested, and which constitutes a well known problem in economy [10].

Since this is the first study enabling agents to participate in each other's reward function, we aimed to establish the empirical groundwork for future research in this direction. For that reason, we took many experimental decisions with simplicity and scalability in mind. However, peer rewards allow for a vast range of novel interactions in MARL, each with the potential of inducing complex behaviours. Hence, we see three main directions in which future studies could push our understanding of peer rewards in reinforcement learning.

The first is to treat peer rewards as a challenge, by designing environments where learning when or how to share rewards is a necessary step towards solving a task. Such settings could be of particular interest for studying emergent communication and automated negotiation between agents, or for testing all works attempting opponent modelling.

The second, as mentioned in our discussion, is to leverage peer rewards as a tool for training and stabilizing multi-agent systems. Indeed, we demonstrated in our experiments that peer rewards can help in promoting a collectively good behaviour even with entirely self-interested agents. We also show that deliberate reward passing – a direct form of peer rewards – can allow agents to acquire skills much faster, likely without requiring any environmental rewards at all. More so, suppose that the peer rewarding mechanic shares characteristics with a target action that we wish our agents would master, as was the case with the range of the gifting and tagging beams in our experiments. If that is the case, one could perform a type of curriculum learning by first allowing agents to interact only through the peer rewarding mechanism, before introducing the target action at a later stage of training. If done correctly, we conjecture that in many cases the agents will learn much faster to use the new action proficiently.

The projected benefit of peer rewards as a tool goes beyond just skill acquisition however. As described at the beginning of the paper, we can imagine agents using rewards as incentives to guide each other towards states of which they are uncertain. They may help coordination, for instance by allowing agents to place incentives, akin to bids or contracts, on desired outcomes. This would also have the added benefit of allowing cross-pollination between multi-agent reinforcement learning and auction theory [4], a subbranch of economics and game theory.

Finally, peer rewards enable us to perform more complex and realistic simulations of real multi-agent systems, such as markets or swarms of insects. In fact, as a side note, it may now be possible to design environments devoid of environmental rewards that still enable agents to learn. This is worth investigating because it improves the applicability of reinforcement learning as a modelling tool in other fields. Additionally, placing reinforcement learning algorithms in novel settings and studying them like a biologist would study an organism can provide insights that a more benchmark-oriented approach would miss. This is in fact part of the approach that we strive for in this paper, with loose inspiration from Perolat et al. [18] and Rahwan et al. [19].

In this line of work, we personally wish to investigate in the near future the effects of deliberate reward passing in settings were "cheaters" are easily identifiable, such as to satisfy all necessary conditions for the emergence of reciprocal altruism. We also aim to study the emergent behaviour of multi-agent systems optimizing exclusively for peer rewards.

REFERENCES

[1] Sherief Abdallah and Michael Kaisers. 2016. Addressing environment non-stationarity by repeating Q-learning updates. *The Journal of Machine Learning Research* 17, 1 (2016), 1582–1612.

[2] Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. 2015. Evolutionary Dynamics of Multi-Agent Learning: A Survey. *Journal of Artificial*

Intelligence Research 53 (08 2015), 659–697. https://doi.org/10.1613/jair.4818

[3] Lucian Bu, Robert Babu, Bart De Schutter, et al. 2008. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38, 2 (2008), 156–172.

[4] Ralph Cassady. 1967. *Auctions and auctioneering*. Univ of California Press.

[5] Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. 2018. Minimalistic Gridworld Environment for OpenAI Gym. https://github.com/maximecb/gym-minigrid. (2018).

[6] Caroline Claus and Craig Boutilier. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI* 1998, 746-752 (1998), 2.

[7] 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.

[8] Jakob Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. 2018. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 122–130.

[9] Jayesh K Gupta, Maxim Egorov, and Mykel Kochenderfer. 2017. Cooperative multiagent control using deep reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems*. Springer, 66–83.

[10] Garrett Hardin. 1968. The tragedy of the commons. *science* 162, 3859 (1968), 1243–1248.

[11] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Çaglar Gülçehre, Pedro A. Ortega, DJ Strouse, Joel Z. Leibo, and Nando de Freitas. 2018. Intrinsic Social Motivation via Causal Influence in Multi-Agent RL. *CoRR* abs/1810.08647 (2018). arXiv:1810.08647 http://arxiv.org/abs/1810.08647

[12] Spiros Kapetanakis and Daniel Kudenko. 2002. Reinforcement learning of coordination in cooperative multi-agent systems. *AAAI/IAAI* 2002 (2002), 326–331.

[13] Martin Lauer and Martin Riedmiller. 2000. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *In Proceedings of the Seventeenth International Conference on Machine Learning*. Citeseer.

[14] 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.

[15] Sanmit Narvekar and Peter Stone. 2019. Learning Curriculum Policies for Reinforcement Learning. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 25–33.

[16] Thanh Thi Nguyen, Ngoc Duy Nguyen, and Saeid Nahavandi. 2018. Deep reinforcement learning for multi-agent systems: a review of challenges, solutions and applications. *arXiv preprint arXiv:1812.11794* (2018).

[17] Georg Ostrovski, Marc G Bellemare, Aäron van den Oord, and Rémi Munos. 2017. Count-based exploration with neural density models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2721–2730.

[18] Julien Perolat, Joel Z Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. 2017. A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems*. 3643–3652.

[19] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. 2019. Machine behaviour. *Nature* 568, 7753 (2019), 477.

[20] Burrhus Frederic Skinner. 1965. *Science and human behavior*. Number 92904. Simon and Schuster.

[21] Christopher Stephens. 1996. Modelling reciprocal altruism. *the British Journal for the Philosophy of Science* 47, 4 (1996), 533–551.

[22] Robert L Trivers. 1971. The evolution of reciprocal altruism. *The Quarterly review of biology* 46, 1 (1971), 35–57.