A Sustainable Ecosystem through Emergent Cooperation in Multi-Agent Reinforcement Learning

Fabian Ritz[1], Daniel Ratke[1], Thomy Phan[1], Lenz Belzner[2] and Claudia Linnhoff-Popien[1]

[1]Mobile and Distributed Systems Group, LMU Munich, Germany
[2]Technische Hochschule Ingolstadt, Germany
fabian.ritz@ifi.lmu.de

Abstract

This paper considers sustainable and cooperative behavior in multi-agent systems. In the proposed predator-prey simulation, multiple selfish predators can learn to act sustainably by maintaining a herd of reproducing prey and further hunt cooperatively for long term benefit. Since the predators face starvation pressure, the scenario can also turn in a tragedy of the commons if selfish individuals decide to greedily hunt down the prey population before their conspecifics do, ultimately leading to extinction of prey and predators. This paper uses Multi-Agent Reinforcement Learning to overcome a collapse of the simulated ecosystem, analyzes the impact factors over multiple dimensions and proposes suitable metrics. We show that up to three predators are able to learn sustainable behavior in form of collective herding under starvation pressure. Complex cooperation in form of group hunting emerges between the predators as their speed is handicapped and the prey is given more degrees of freedom to escape. The implementation of environment and reinforcement learning pipeline is available online.[1]

Introduction

Sustainable ecosystem management offers methods to prevent natural resources from being over-exploited so that the utility of a given ecosystem is maintained for a longer period of time. E.g., Smith et al. (2016) survey indicators for sustainable ecosystem management and different impact stages that ecosystems can be at in terms of exploitation. In this paper, a predator-prey simulation models an artificial ecosystem. If the predators act sustainably by not catching the prey faster than it reproduces, they can prevent the ecosystem from collapse while not starving themselves.

Reinforcement Learning (*RL*) has been increasingly proposed to solve optimization problems such as plant development for sustainable agriculture (Binas et al., 2019) or crop yield prediction (Elavarasan and Vincent, 2020). An extensive survey on using RL for sustainable energy systems has been done by Yang et al. (2020), showing a broad range of applications such as integrating renewable energy with their implied uncertainty into energy networks while optimizing economic objectives, energy utilization and environmental impacts. Recently, Ritz et al. (2020) analyzed a predator-prey environment in terms of sustainable behavior. They showed that a single RL predator can learn to maintain a herd of prey in absence of starvation for long-term benefit.

This paper extends the scenario of Ritz et al. (2020) to multiple RL predators and further adds starvation, putting pressure on the predators. According to game theory, this might turn into a *tragedy of the commons* if selfish individuals decide to greedily hunt down the prey population before their conspecifics do, ultimately leading to extinction prey and predators. While such problems are studied as *iterated Prisoner's Dilemmas* in the field of (evolutionary) game theory (Axelrod and Dion, 1988) and as *Sequential Social Dilemmas* in the field of Multi-Agent Reinforcement Learning (*MARL*) (Leibo et al., 2017; Pérolat et al., 2017), most research is limited to repeated games or discrete environments and no approach considered sustainability under starvation pressure so far. However, if the predators learn collective herding, which may be seen as a form of cooperation, it will be interesting to analyze whether additional cooperation emerges on top of herding such as group hunting. Back in 2005, the Science Magazine placed the question of how cooperative behavior evolved within the top 25 open problems of science (Pennisi, 2005) and up to today, complex cooperation remains an open field of research within the AI community (Dafoe et al., 2020).

This paper jointly analyzes impact factors inducing *sustainable and cooperative* behavior among up to three selfish RL predators over multiple dimensions and further proposes suitable metrics. Despite starvation pressure, MARL is found to overcome a collapse of the simulated ecosystem. Even though we refrain from handcrafted reward shaping to avoid inducing a bias, we find the agents to succeed if the environment provides suitable learning conditions. While our immediate goal is to train sustainable, cooperative agents and to define best practices on how to achieve this, we hope that further down the road, the powerful toolbox of RL can be used to analyze current practices across other domains and help build a sustainable future by optimizing those.

[1]Code available at: https://github.com/instance01/fish-rl-alife

Foundations and Related Work

Reinforcement Learning

In Reinforcement Learning (*RL*), an agent observes an environment and acts upon it based on learned rules. Formally, the environment can be described as a Markov Decision Process (*MDP*), which consists of a tuple (S, A, R, P), where S is the set of states, A is the set of actions and R is the reward function. Reward r_t is given after executing an action a_t in a given state s_t at time step t and P is the transition probability matrix which defines the probability of ending up in state s_{t+1} after executing a_t in s_t. The goal of the agent is to learn a policy $\pi(a_t|s_t)$ that maximizes the total expected reward $\sum_{i=0}^{T-1} \gamma^i r_{t+i}$ when executing $a_t \sim \pi(a_t|s_t)$, where $\gamma \in [0, 1]$ is the discount factor and T is the horizon. In this paper, the state is partially observable, which is formalized by a partially observable MDP (*POMDP*). A POMDP additionally contains the set of local observations O, and the set of observation probabilities Ω. Although value-based RL approaches like Deep Q-learning are popular to learn a policy π (Mnih et al., 2015), this paper uses Proximal Policy Optimization (*PPO*), a policy gradient method that recently showed impressive results in many benchmark environments (Schulman et al., 2017).

Multi-Agent Reinforcement Learning

Multi-Agent Reinforcement Learning (*MARL*) is RL with n agents in a shared environment. Therefore, $O = (O_k)_{1 \leq k \leq n}$ is the joint observation space with O_k as the observation space of agent k and $A = (A_k)_{1 \leq k \leq n}$ is the joint action space with A_k as the action space of agent k. P and the joint reward $r_t \in R$ are based on the joint observation $o_t \in O$ and the joint action $a_t \in A$, and each agent maximizes its own expected reward of o_t and a_t. State-of-the-art MARL algorithms such as QMIX (Rashid et al., 2018) exploit global information like joint observations and joint actions in a centralized training regime in order to learn individual policies that can be executed decentralized. A more scalable alternative is Independent Learning (*IL*), i.e. to use single-agent RL algorithms and let each of the n agents learn on its own. However, this comes at the cost of non-stationary dynamics as for a given agent, all others interfere with the environment. This violates the Markov property as the transition from one state to another also depends on the actions of all other agents. Thus, there is no guarantee of convergence and achieving stable policies is generally more difficult. E.g., Laurent et al. (2011) show how in some cases one agent exploring the environment may invalidate a good policy of another agent. Still, Sunehag et al. (2019) find the emergence of advanced strategies emerge by independent MARL in simulated multiple-species ecosystems. In accordance to that, de Witt et al. (2020) also find that PPO is somewhat robust to non-stationarity and thus a suitable candidate for IL.

Herding

One of nature's most remarkable relationships is the predator-prey dynamic, in which predators and prey go through shifted cycles of high and low population density. Much work has been done to model and empirically show these dynamics, e.g. via the *Lotka-Volterra model*. E.g., Blasius et al. (2020) set forth that predators and their prey can sustainably co-exist in these cycles for a very long time. While predators do exploit at times and end up depleting the prey population, they never over-exploit it irresistibly though. Using MARL, Yang et al. (2018) study population dynamics with predators learning greedily by maximizing their individual reward in a setting that includes predator starvation. Also, predators may hunt in groups to increase the chance of a successful catch but also share the reward afterwards. They find that the cyclic dynamics described by the Lotka-Volterra model emerge. Further, Wallach et al. (2015) propose that apex predators self-regulate their own population to make sure that there is no over-exploitation, i.e. keep their own population at a certain level, which limits the pressure on the prey population.

Regarding self-regulation, prior work of Ritz et al. (2020) only considered a single predator. This paper assumes a group of non reproducing predators under starvation pressure and analyzes if collective self-regulation, i.e. choosing not to hunt temporary but waiting for a prey population to recover, emerges. We call this *herding*, i.e. maintaining a herd of prey and never over-exploiting while hunting freely if the herd size allows so.

Cooperation

Complex cooperation, e.g. in the form of group hunting, can also be found in nature. For instance, lions cooperate, even though it is in a selfish way. Amongst others, Scheel and Packer (1991) show that they often 'refrain' when easy prey such as warthogs are hunted while joining the hunt when difficult prey such as zebras are hunted. Refraining from the hunt still enables them to join the feast after the prey is caught, but with less total energy used for the hunt. The authors compare this to cheating. Computational analysis by Burtsev (2005) found that resource supply influences the rate of peaceful cooperation versus aggression. Further, Tanabe and Masuda (2012) argue that RL is able to learn cooperation through natural selection, showing the *Baldwin effect*. Using independent MARL in simulated grids, Leibo et al. (2017) analyze how conflicts can emerge from competition over shared resources. They observe that two learning agents collecting resources act more aggressive when resources are scarce and that for cooperation between two predators in predator-prey interaction to arise, additional rewards for cooperative catches have a significant impact. Scaling the scenario of Leibo et al. (2017) to a common pool resource appropriation problem, Pérolat et al. (2017) also introduce social outcome metrics including sustainability.

Assuming predators are generally selfish, this paper aims for cooperative behavior beyond overcoming the tragedy of the commons in form of group hunting and analyzes which factors impact such cooperation. Prior work (Leibo et al., 2017; Pérolat et al., 2017; Ritz et al., 2020) does not consider predator starvation, which we additionally regard in this paper. Apart from splitting the total reward for cooperatively caught prey, there is no reward shaping. Finally, we contribute experiments in a three-agent setting which are considered significantly more difficult from a game-theoretic- and an RL-standpoint compared to a two-agent setting.

Further Emergent Behavior

Inspired by nature, emergent behavior has been studied using computer models under further aspects. E.g., Olson et al. (2016) analyze predator-prey interaction in the context of *swarming* using an evolutionary algorithm. They find that the swarming among prey is influenced by how predators attack. While Reynolds (1999) proposes static rules leading to swarming, Morihiro et al. (2008) have their agents learn to do so by encoding these rules in the reward of an RL algorithm. Further, Sunehag et al. (2019) find the emergence of flocking and symbiosis with rewards shaping by independent MARL in simulated multiple-species ecosystems. In parallel, Hahn et al. (2019) demonstrate the emergence of swarming without reward shaping and later show that in their scenario, prey swarming is a Nash Equilibrium (Hahn et al., 2020) and the prey could perform better if collectively *fleeing independently*. Ritz et al. (2020) also found independently fleeing prey to be significantly harder to hunt for a single RL predator. Based on preliminary experiments (see Fig.2a), this paper also uses independently fleeing prey.

Yet, emergent behavior can also be observed in non predator-prey scenarios. E.g., Leibo et al. (2019) study an evolutionary, population based approach and find the emergence of division of labor. While that is out of scope for this paper, our results might generalize to such scenarios as well.

Domain

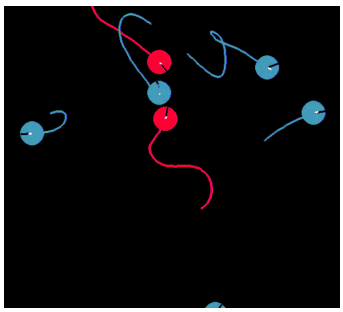


Figure 1: Rendering of the simulated ecosystem. Two RL predators (red) cooperatively isolate and catch a prey (blue). The corresponding tails indicate the moving trajectories.

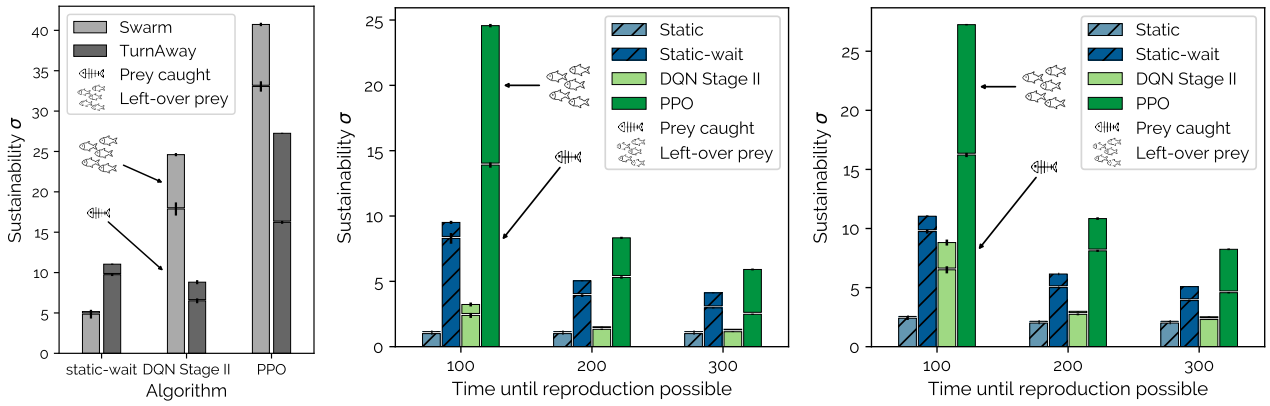
We study sustainable and cooperative behavior in a continuous, two-dimensional simulation with two types of agents, predator (e.g. sharks) and prey (e.g. small fish). Both are represented by unicycles, a model of mobile robotics. Depending on the respective scenario, the environment is either bounded by walls or open, forming a torus, i.e. all borders wrap around. All agents move by adjusting their linear velocity (acceleration) and their angular velocity (orientation), modeling double integrator dynamics. Maximum speed is constrained via simulated friction, allowing flexible agent setups. By default, the view distance of predators is greater than the world size, while the view distance of the prey is restricted to a fraction of the world size. When agents overlap with each other or the walls, an elastic collision is performed. If a predator collides with a conspecific while facing it, the conspecific is stunned, i.e. it floats for a number of steps. If both both predators face each other while colliding, both are stunned. If a predator collides with a prey, the prey is considered caught and removed from the simulation. Prey may reproduce, i.e. spawn a conspecific at the current position, if all following conditions are met: No predator is within the view radius of the prey, a certain number of time steps since the last reproduction has passed and the prey population limit is not exceeded. Lastly, predators that do not catch enough prey can die from starvation. Their initial survival time varies throughout the experiments while the additional survival time per caught prey is fixed.

Actions

The action space A_k is a triple with two continuous values representing acceleration, orientation and a boolean whether to reproduce. Prey always reproduce as soon as possible and predators do not reproduce. Acceleration is clamped to $[-1, 1]$ and orientation is clamped to $[-180°, 180°]$.

Observations

The observation space O_k is structured in a uniform way for predator and prey. It includes the current orientation of the agent, the readiness to reproduce, a list of walls and a list of other agents. Depending on view radius and limits explained in the following, only a subset of all walls and agents can be observed, making the environment partially observable. The list of walls consists of distance and angle from the current agent to each of the closest n_w walls. If no walls are perceivable, e.g. due to a limited view radius or because the environment is a torus, zero entries maintain a consistent observation space. Next, n_pred triple slots are used for predators. Each triple slot consists of distance, angle and orientation of the predator. Again, predators outside of the view radius lead to zero entries. Finally, n_prey slots of the same structure are available for the prey. As the three aforementioned parameters remaining static throughout training and evaluation. The zero-padding ensures a fixed-size observation independent of the perceivable entities.



(a) Sustainability of swarming and independently fleeing prey.

(b) Sustainability with one initial prey.

(c) Sustainability with two initial prey.

Figure 2: Average sustainability σ (sum of caught and left-over prey at the end of the episode) of different predator algorithms. PPO consistently achieves the most sustainable behavior due to herding a rather large prey population, resulting in a high rate of prey reproduction and thus more prey to be caught.

Rewards

RL predators are granted a reward of 10 for each caught prey. In settings with multiple predators, this reward can be split up: If n predators are within the *shared catch zone*, i.e. within a defined radius from the place of capture, each predator gets a reward of $\frac{10}{n}$. There are no other (shaped) rewards, avoiding the induction of further bias.

Experimental Setup

Table 1: PPO Hyperparameters

total domain steps	21×10^6
entropy coefficient	0.0
learning rate	9×10^{-5}
value function coefficient	0.5
max gradient norm	0.5
γ	0.99
λ	0.95
mini-batches	4
optimizer epochs	4
clipping range	0.1
hidden layer neurons	3×64

To assess whether the RL predators learn sustainable and cooperative behavior, the following metrics are used: First, **sustainability** σ is defined the sum of caught and the left-over prey at the end of an episode. Maximum sustainability would entail catching a lot of prey while allowing a large population to live. To ensure significance, σ is gathered and averaged over multiple episodes. Further, (collective) herding requires all RL predators to slow down and wait or slowly stalk behind prey when the prey population is low while hunting actively at high speed when the prey population rises again[2]. Specifically, herding does not include strolling around continuously and hunting prey nearby by chance. While such behavior is difficult to capture mathematically, its results can be accessed through the **herding ratio** η which is gathered over multiple episodes: Per episode, herding is considered successful if one to ten prey agents are alive (if the prey population hits the limit of eleven, we consider the predators unable to hunt, thus unable to herd) and unsuccessful otherwise. η is defined as the number of the episodes that resulted in herding divided by the total number of episodes. Next, the **failure ratio** ϕ, defined as $1 - \eta$, expresses how often the RL predators failed to either maintain a herd of prey or to catch any prey at all. Finally, the **cooperation ratio** κ assesses group hunting among RL predators. A catch is considered cooperative if all RL predators are within the shared catch zone and lone otherwise. κ is defined the number of cooperative catches divided by the total number of catches. Similar to σ, κ is gathered and averaged over multiple episodes.

Regarding the agents, the prey strategies proposed by Hahn et al. (2019) were compared in preliminary experiments (c.f. Figure 2a). In accordance with previous work (Ritz et al., 2020), independently fleeing was used for all following experiments as it was found significantly harder to hunt than swarming prey. The RL predators were trained with PPO. The most important hyperparameters are shown in Table 1. The neural network used by PPO is set up as three fully connected layers á 64 nodes with batch normalization. We modified the implementation of Dhariwal

[2]Illustrative videos can be found here: shorturl.at/jnoKO

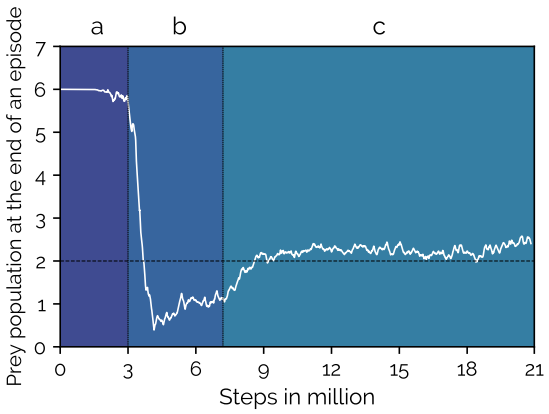


Figure 3: Left-over prey population per learning episode over 21M steps. In this demonstration setting, two predators hunt at most six prey. As the predators learn to herd prey, they go through three stages. In the first stage a, they cannot hunt effectively yet and thus leave the full population intact until the end of the episode. In the following stage b, they effectively hunt and over-exploit, leaving no prey at the end of the episode. Lastly, in stage c, they learn to maintain a herd of approx. two prey until the end of the episode.

et al. (2017) to support for multi-agent independent RL. To ensure comparability with prior work of Ritz et al. (2020), we benchmark a single PPO agent against their best performing agent, *DQN Stage II*. We reproduced their training pipeline and applied the hyperparameters reported to result in highest average number of caught prey per episode. Further, two non-RL agents are provided as baselines: The *static* algorithm chases the nearest prey based on heuristics and the *static-wait* algorithm does the same except when there is only one prey left. At that point, the predator idles, giving the prey a chance to escape and reproduce.

Regarding training and evaluation, each episode lasts 3000 domain steps. PPO is trained for 7000 episodes, i.e. in 21×10^6 domain steps, with constant hyperparameters. If not stated differently, the initial survival time of predators is 3000 steps, i.e. starvation is impossible, and stunning is disabled. Further environment settings remain at the default (details in the source code) if not specified otherwise. Every result is the average of 400 values gathered from 20 independently trained models, each evaluated 20 times.

Results

To bridge the gap to previous work of Ritz et al. (2020), Figure 2b and Figure 2c compare the performance of a single predator in two scenarios with a low amount of initial prey. The predators need to be able to restrain their greediness and keep prey alive to allow for a larger population to grow and thus the overall reward to increase. Here, the RL algorithm

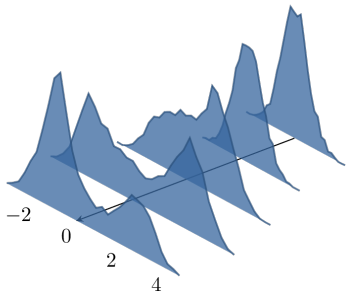


Figure 4: From back to front: Distribution of predator acceleration over time. As the predators learn to herd prey during the training, the distribution becomes bimodal. This suggests that in some cases, they deliberately wait (for the prey to reproduce), while in other cases they move fast (hunting).

PPO achieves the highest sustainability. Except one case, PPO manages to hunt the most prey while keeping a sufficient prey population. The RL algorithm DQN Stage II is not able to compete. The two deterministic algorithms, static and static-wait, perform as intended and are unable to beat PPO, keeping none or one prey alive per design. A tangential improvement of PPO against DQN Stage II is the easier training pipeline: While DQN needed a handcrafted curriculum with two stages, PPO is trained end-to-end without additional handcrafted adaptations to the training process.

Moving on to emergent herding, Figure 3 illustrates the learning process of two RL predators going through three stages: (a) Being unable to hunt, (b) hunting greedily while over-exploiting, and (c) hunting in a sustainable manner. An indicator for herding behavior is given by Figure 4, showing that during training, the distribution of the predators' acceleration gradually turns from a unimodal one (always moving at full speed) to a bimodal one (partially waiting, partially moving). Figure 5 shows two major impact factors for herding to emerge. Firstly, the number of initial prey is crucial for the herding ratio η. This may be due to the different amount of time until the RL predators get into a situation with few or none prey left. It takes long to hunt down a large initial prey population and a lot of reward can be collected until the prey is extincted. Contrary, a small initial prey population is depleted faster and the trade-off between immediate and future rewards comes into play sooner. Secondly, starvation pressure plays an integral part in herding. Depending on the amount of initial prey and initial predator survival time, the sweet spot with the highest η varies. Yet, the most concise results can be observed with few initial prey. In such scenarios, predators with few initial survival time do not act sustainably due to the immediate fight for survival, prohibiting herding. Contrary, granting the predators much initial survival time lowers the need for hunting.

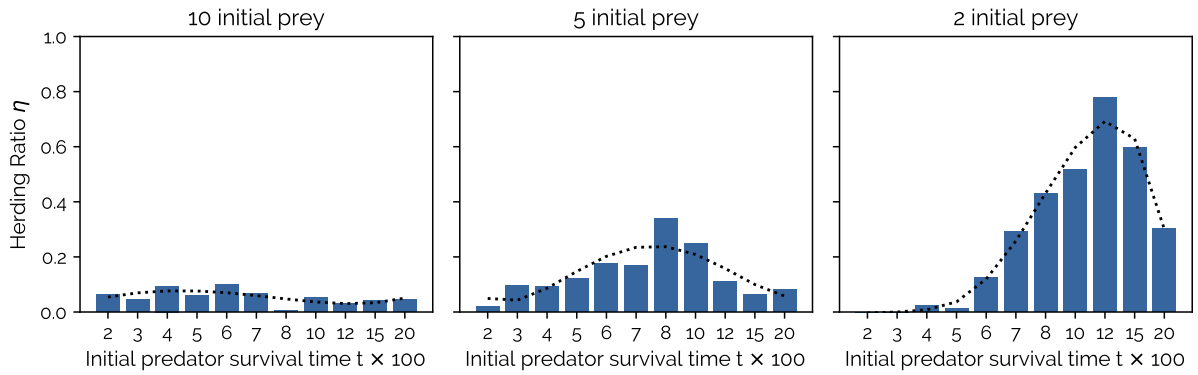
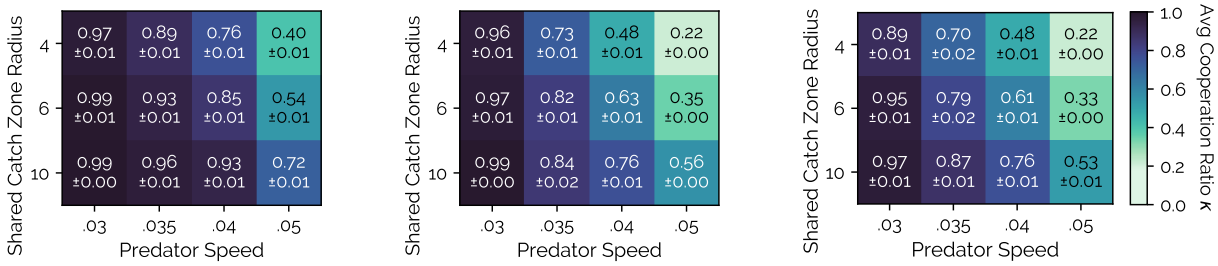
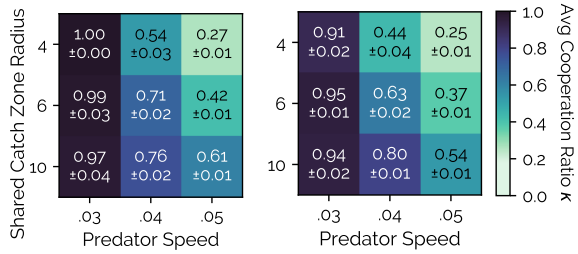


Figure 5: Impact factors on sustainability of two RL predators measured via the herding ratio η. From left to right, the initial prey population is decreased. A lower initial prey population increases the survival pressure. Further, each scenario varies the initial predator survival time. Given few initial prey, high initial survival time results in the highest η. Yet, combining few initial prey with few initial survival time induces so much survival pressure that no herding can be observed. The curve was fitted with a fourth-degree polynomial.



(a) κ with 5 initial prey. (b) κ with 10 initial prey. (c) κ with 10 initial prey and stuns.

Figure 6: Impact factors on group hunting of two RL predators measured via the average cooperation ratio κ. From left to right, the initial prey population is increased and predator stun mechanics are added. Each scenario varies the shared catch zone radius and the predators' speed. Overall, κ increases if there is fewer prey, if the predators are slower and if the the shared catch zone radius is larger. For reference, the maximum speed of prey is 0.08.



(a) 500 init. survival time. (b) 1500 initial survival time.

Figure 7: Average Cooperation ratio κ of two RL predators with ten initial prey and varying predator starvation pressure. Each scenario varies the shared catch zone radius and the predators' speed. Due to the high number of initial prey, the differences in the cooperation rate are negligible.

After achieving sustainable behavior through emergent herding, complex cooperation in form of group hunting remains to be shown. Handcrafted reward shaping could induce a bias and shall be avoided in this paper, so only the environmental setting is left to incentivize the predators to cooperate. Accordingly, the maximum speed of the predators is lowered so that catching up with the prey takes longer. Further, the walls bounding the environment are replaced with a wrap-around, allowing more degrees of freedom for the prey to escape and avoiding that a single predator catches prey alone easily by driving it into a corner. Lastly, the shared catch zone is added. Assuming that all predators within a certain radius effectively contributed to the hunt and may benefit from the catch, this zone is used both to measure cooperation and to split up the reward between the involved predators. Overall, the expected behavior for the predators in the following scenarios is to cooperate to catch prey.

As a first glance of success, Figure 1 shows two predators surrounding a prey from two sides and catching it. Analyzing the details, Figure 6 demonstrates how larger shared catch zone radii and slower predator speeds lead to a higher cooperation rate κ. Similar to herding, less initial prey leads to higher κ due to increased environment difficulty. With less prey, there are less (accidental) lone catches due to overcrowding and the prey evade easier since there are less obstacles (other prey) to take into account.

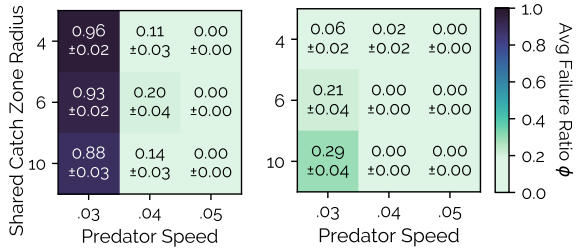
However, not all environmental factors actually influence cooperation. Stuns (c.f. Figure 6c) have no major impact. Predators learn to not crash into each other, since a stunned partner leads to less catches, but this does not impact the overall learning performance. Further experiments were carried out to analyze stunning behavior. The hypothesis was that, similar to Leibo et al. (2017), RL predators use stuns more if less prey is available, such that the stunned predator cannot steal prey from the prey population. However, no significant correlation between the stun rate and the number of prey could be observed.

As to be seen in Figure 7, adding starvation pressure surprisingly does not induce greedy competition. However, Figure 8 points out that slow moving RL predators, which so far were most cooperative, face high failure ratios ϕ. Learning seems to be too inhibited: An initial predator survival time of 500 time steps gives the predators very few time to learn how to catch prey. Yet, the few remaining catches do happen in cooperation (c.f. Figure 7). It should be mentioned that ϕ is very small in absence of starvation pressure.

Another major impact factor to cooperation is the view distance of prey. The further it can see, the faster it can react to and evade predators. Doubling the view distance from 10 to 20 results in (almost) full cooperation in 9 out of 12 scenarios (c.f. Figure 9).

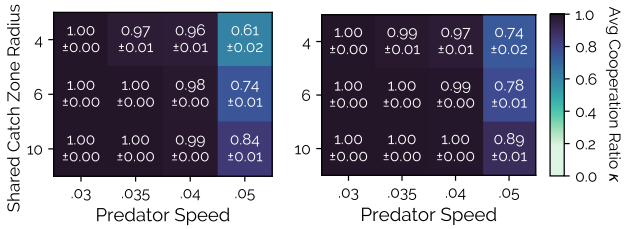
Finally, the experiments are scaled to three RL predators. Within all tested prey view distances, 25 yields the highest cooperation ratio κ. Figure 10 shows the average κ over 20 models, each evaluated 100 times. While all of the prey view distances nearly reach full cooperation between two predators, even the most favorable setting does never cause more than $\kappa = 0.59$ in the three predator setting. Further increasing the view distance or slowing the predators down lead to impracticably high failure ratios ϕ.

So far, all experiments were performed using Independent Learning (*IL*). To outline scalability for future work, a comparison between Parameter Sharing (*PS*) and IL was done. In PS, all agents share the neural network parameters. Figure 10 shows that PS causes less cooperation while Figure 11 shows that the average number of caught prey stays the same. This suggests that PS has better performance as such agents are able to catch prey with less predators involved. Additionally, PS finishes training on average 2.63 times faster (18:56 hours $\pm$ 00:05 for PS versus 50:08 hours $\pm$ 00:40 for IL) than IL in the experiments discussed here.



(a) 500 init. survival time. (b) 1500 initial survival time.

Figure 8: Average Failure ratio ϕ of two RL predators with varying predator starvation pressure. Each scenario varies the shared catch zone radius and the predators' speed. Figure 8a shows that few initial survival time results in high ϕ. While the remaining few catches (c.f. Figure 7a) are cooperatively, low predator speed combined with few initial survival time results in few predators being able to catch even a single prey. Thus, starvation pressure may inhibit learning.

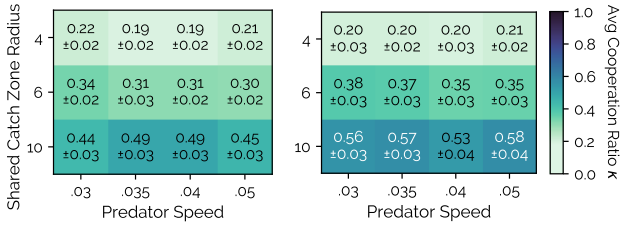


(a) Prey view distance 20. (b) Prey view distance 30.

Figure 9: Average cooperation rate κ of two RL predators with varying values for prey view distance, shared catch zone radius and predator speed. The further prey see, the tougher it is to catch them. Doubling the default of 10 (see previous figures) has more impact than increasing it further.

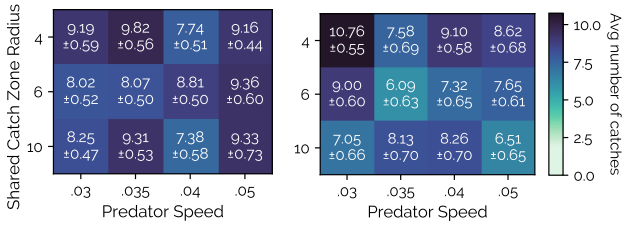
Discussion

Theoretically, RL algorithms result in solely rational behavior maximizing the given optimization problem. In practice, RL agents often converge to less efficient equilibria, especially in Multi-Agent settings. In our scenario, a tragedy of the commons would have occurred if any of the agents caused over-exploitation. However, this was not the case. We hypothesize that when an RL predator observes the other(s) catching the last prey and thus preventing any future rewards, it is equally as important as if it did so itself. A common rule learned by multiple predators at the same time might have been: If the number of remaining prey is smaller than x, death is imminent, which should be avoided. Therefore, we expect that current state-of-the-art RL is able to learn equivalents of the rules that evolved in the social structures of animals and humans which can avert a tragedy of



(a) Parameter Sharing (PS) (b) Independent Learning (IL)

Figure 10: Average cooperation ratio κ of three RL predators with different MARL architectures. Catches are only considered cooperative if all predators are involved. Each scenario varies the shared catch zone radius and the predators' speed. Parameter Sharing results in slightly lower cooperation ratio κ than IL.



(a) Parameter Sharing (PS). (b) Independent Learning (IL).

Figure 11: Average number of catches of three RL predators with different MARL architectures. Each scenario varies the shared catch zone radius and the predators' speed.

the commons. However, the results also suggest that for RL to be sustainable, the scenario must provide suitable learning conditions such as the ratio of predator and prey speed or the prey view distance: The herding ratio κ was significant only under certain environmental settings.

Further, we argue that overcoming the tragedy of the commons via (collective) herding may be cooperation from a game-theoretic standpoint already, but can be achieved without reasoning about other agents. Therefore, herding is only as one step towards more complex cooperation in form of group hunting. Initially, predators restrained themselves to not catch prey too quickly and thus followed a common goal, but also hunted in vigorous competition when the population was large enough. This was observed especially when the prey population raised just above the threshold at which predators start hunting. Nevertheless, group hunting among the RL predators emerged on top of herding once the environment settings were adjusted so that solely cooperative catches yielded high rewards and guaranteed survival. It is interesting that these findings share similarities with Axelrod and Dion (1988), summarizing impact factors for cooperation based on reciprocity to arise, although we use MARL instead of evolutionary algorithms. While the emergence of cooperation in a setting with no risk of being exploited might be trivial from a game-theoretic standpoint, we argue that considering Dafoe et al. (2020), successful complex cooperation is noteworthy as our RL predators can neither coordinate their actions directly, e.g. through communication, nor have central institutions, e.g. social norms or rules, but learn and act based on past observations only. Yet, scaling the experiments to three agents showed that even with current state-of-the-art RL, this is a boundary walk between making the task too easy, thus allowing agents to individually succeed, and making the task too difficult, thus overstraining the (computational) capabilities of the RL agents.

Future work might consider to scale the scenario further with Parameter Sharing (*PS*) as from a biological perspective, it may be compared to observational learning where an animal copies behavior it has seen from another animal: Rodríguez et al. (2014) posit that a special type of neurons in the brain, the *mirror neuron*, is responsible for observational learning. In MARL, PS would be an extreme form of observational learning, where each agent copies the action another agent took.

Conclusion

So far, it was known that a single RL predator may learn herding of prey (Ritz et al., 2020). This paper applied MARL to a predator-prey scenario and showed that two selfish predators are able to learn sustainable behavior in form of collective herding even under starvation pressure. Naturally, starvation pressure heavily impacts the herding ratio. Further, complex cooperation in form of group hunting emerged between the selfish predators as their speed was handicapped and the prey was given more degrees of freedom to escape. Lastly, experiments were successfully scaled to three RL predators. This suggests that MARL can be used for problems requiring sustainable and cooperative behavior if suitable learning conditions are provided.

Future work might consider self-replicating predators, scaling the scenario to more than three predators and mixing of agents trained under different environmental conditions to answer the following questions: How many predators can a given environment feed? Can the predators learn to regulate their population accordingly? Do some of them learn to collude and let others die off? What impact do mixed teams of differently trained agents have? Lastly, it would be interesting to apply *Learning with Opponent Learning Awareness* by Foerster et al. (2018). Such predators could learn to predict what the others are currently learning and adapt their behavior accordingly.

References

Axelrod, R. and Dion, D. (1988). The further evolution of cooperation. *Science*, 242(4884):1385–1390.

Binas, J., Luginbuehl, L., and Bengio, Y. (2019). Reinforcement learning for sustainable agriculture. In *ICML 2019 Workshop Climate Change: How Can AI Help*.

Blasius, B., Rudolf, L., Weithoff, G., Gaedke, U., and Fussmann, G. F. (2020). Long-term cyclic persistence in an experimental predator–prey system. *Nature*, 577(7789):226–230.

Burtsev, M. S. (2005). Artificial life meets anthropology: A case of aggression in primitive societies. In *European Conference on Artificial Life*, pages 655–664.

Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., Larson, K., and Graepel, T. (2020). Open problems in cooperative ai. *arXiv preprint arxiv:2012.08630*.

de Witt, C. S., Gupta, T., Makoviichuk, D., Makoviychuk, V., Torr, P. H., Sun, M., and Whiteson, S. (2020). Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*.

Dhariwal, P., Hesse, C., Klimov, O., Nichol, A., Plappert, M., Radford, A., Schulman, J., Sidor, S., Wu, Y., and Zhokhov, P. (2017). Openai baselines. `https://github.com/openai/baselines`.

Elavarasan, D. and Vincent, P. D. (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. *IEEE Access*, 8:86886–86901.

Foerster, J., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. (2018). Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, page 122–130.

Hahn, C., Phan, T., Feld, S., Roch, C., Ritz, F., Sedlmeier, A., Gabor, T., and Linnhoff-Popien, C. (2020). Nash equilibria in multi-agent swarms. In *12th International Conference on Agents and Artificial Intelligence (ICAART 2020)*.

Hahn, C., Phan, T., Gabor, T., Belzner, L., and Linnhoff-Popien, C. (2019). Emergent escape-based flocking behavior using multi-agent reinforcement learning. In *Artificial Life Conference Proceedings*, pages 598–605.

Laurent, G. J., Matignon, L., Fort-Piat, L., et al. (2011). The world of independent learners is not markovian. *International Journal of Knowledge-based and Intelligent Engineering Systems*, 15(1):55–64.

Leibo, J. Z., Perolat, J., Hughes, E., Wheelwright, S., Marblestone, A. H., Duéñez Guzmán, E., Sunehag, P., Dunning, I., and Graepel, T. (2019). Malthusian reinforcement learning. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, page 1099–1107.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, page 464–473.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Morihiro, K., Nishimura, H., Isokawa, T., and Matsui, N. (2008). Learning grouping and anti-predator behaviors for multi-agent systems. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 426–433.

Olson, R. S., Knoester, D. B., and Adami, C. (2016). Evolution of swarming behavior is shaped by how predators attack. *Artificial Life*, 22(3):299–318.

Pennisi, E. (2005). How did cooperative behavior evolve? *Science*, 309(5731):93–93.

Pérolat, J., Leibo, J. Z., Zambaldi, V., Beattie, C., Tuyls, K., and Graepel, T. (2017). A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems*, volume 30.

Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., and Whiteson, S. (2018). Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 4295–4304.

Reynolds, C. W. (1999). Steering behaviors for autonomous characters.

Ritz, F., Hohnstein, F., Müller, R., Phan, T., Gabor, T., Hahn, C., and Linnhoff-Popien, C. (2020). Towards ecosystem management from greedy reinforcement learning in a predator-prey setting. In *Artificial Life Conference Proceedings*, pages 518–525.

Rodríguez, Á. L., Cheeran, B., Koch, G., Hortobágyi, T., and Fernandez-del Olmo, M. (2014). The role of mirror neurons in observational motor learning: an integrative review. *European Journal of Human Movement*, (32):82–103.

Scheel, D. and Packer, C. (1991). Group hunting behaviour of lions: a search for cooperation. *Animal behaviour*, 41(4):697–709.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Smith, A., Berry, P., and Harrison, P. (2016). Sustainable ecosystem management. *OpenNESS Ecosystem Services Reference Book. EC FP7 Grant Agreement*, (308428).

Sunehag, P., Lever, G., Liu, S., Merel, J., Heess, N., Leibo, J. Z., Hughes, E., Eccles, T., and Graepel, T. (2019). Reinforcement learning agents acquire flocking and symbiotic behaviour in simulated ecosystems. In *Artificial Life Conference Proceedings*, pages 103–110.

Tanabe, S. and Masuda, N. (2012). Evolution of cooperation facilitated by reinforcement learning with adaptive aspiration levels. *Journal of theoretical biology*, 293:151–160.

Wallach, A. D., Izhaki, I., Toms, J. D., Ripple, W. J., and Shanas, U. (2015). What is an apex predator? *Oikos*, 124(11):1453–1461.

Yang, T., Zhao, L., Li, W., and Zomaya, A. Y. (2020). Reinforcement learning in sustainable energy and electric systems: A survey. *Annual Reviews in Control*.

Yang, Y., Yu, L., Bai, Y., Wen, Y., Zhang, W., and Wang, J. (2018). A study of ai population dynamics with million-agent reinforcement learning. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 2133–2135.