Reinforcement Learning Agents acquire Flocking and Symbiotic Behaviour in Simulated Ecosystems

Peter Sunehag, Guy Lever, Siqi Liu, Josh Merel, Nicolas Heess,
Joel Z. Leibo, Edward Hughes, Tom Eccles,Thore Graepel
DeepMind, London UK, sunehag@google.com

Abstract

In nature, group behaviours such as flocking as well as cross-species symbiotic partnerships are observed in vastly different forms and circumstances. We hypothesize that such strategies can arise in response to generic predator-prey pressures in a spatial environment with range-limited sensation and action. We evaluate whether these forms of coordination can emerge by independent multi-agent reinforcement learning in simple multiple-species ecosystems. In contrast to prior work, we avoid hand-crafted shaping rewards, specific actions, or dynamics that would directly encourage coordination across agents. Instead we test whether coordination emerges as a consequence of adaptation without encouraging these specific forms of coordination, which only has indirect benefit. Our simulated ecosystems consist of a generic food chain involving three trophic levels: apex predator, mid-level predator, and prey. We conduct experiments on two different platforms, a 3D physics engine with tens of agents as well as in a 2D grid world with up to thousands. The results clearly confirm our hypothesis and show substantial coordination both within and across species. To obtain these results, we leverage and adapt recent advances in deep reinforcement learning within an ecosystem training protocol featuring homogeneous groups of independent agents from different species (sets of policies), acting in many different random combinations in parallel habitats. The policies utilize neural network architectures that are invariant to agent individuality but not type (species) and that generalize across varying numbers of observed other agents. While the emergence of complexity in artificial ecosystems have long been studied in the artificial life community, the focus has been more on individual complexity and genetic algorithms or explicit modelling, and less on group complexity and reinforcement learning emphasized in this article. Unlike what the name and intuition suggests, reinforcement learning adapts over evolutionary history rather than a life-time and is here addressing the sequential optimization of fitness that is usually approached by genetic algorithms in the artificial life community. We utilize a shift from procedures to objectives, allowing us to bring new powerful machinery to bare, and we see emergence of complex behaviour from a sequence of simple optimization problems.

Introduction

Our natural world is the ultimate example of a self-organizing system (Ashby, 1947). Species and individuals adapt to each other in competition and cooperation, often as predators and prey in food chains. One ubiquitous example of cooperative group behavior is flocking, which can be found on land, sea and air, and numerous benefits from flocking for both predators and prey have been discussed in the literature (Handegard et al., 2012; Ruxton, 2012). For instance, if predators are sparse, a flocked group of prey is not much more likely to be detected than any single individual. Thus, if the predator eliminates at most one individual per detection, it follows that fewer prey will be eaten if they stick together. Further, if the prey are collectively more likely to detect the predator and thus to avoid predation, this improves individual survival chances. Flocking is not only used by prey species but also by predators: it can enable predators to cut off escape routes for a group of prey (e.g. seatrout hunting juvenile gulf menhaden) (Handegard et al., 2012); enable species to jointly capture larger prey, e.g. humans hunting whales (Alvard, 2003); or reduce individual nutritional variability by sharing captures. A second example of group behavior are symbiotic partnerships between species, for example humming bird nests are safer from jay predation when a hawk, which threatens the jays, is situated on top of the same tree (Greeney et al., 2015).

We hypothesize that group strategies like flocking and symbiosis can result in response to very generic predator-prey pressures and opportunities in a spatial environment with range-limited sensation and action, and we test this hypothesis experimentally by deploying independent reinforcement learning (RL) agents in generic simulated ecosystems. RL agents, like e.g. genetic algorithms, learn across the full (evolutionary) history and not primarily during episodes (life-times) and is primarily here viewed as a powerful way to optimizing the sequence of optimization problems posed by the ecosystem including the policies of the other species at the relevant times.

Our environments have three trophic levels (prey, predators and apex predators) and thus enable the emergence of partnerships within and across species. Further, unlike prior work (Morihiro et al., 2006; Hung, 2015; Yang et al., 2018) we do not shape the dynamics, actions or rewards to specifically encourage or facilitate particular group behaviors, Instead we show that such behaviors can emerge in simple general contexts. We use two simulation platforms for this study, to highlight the generality of the findings and show that implementation details are not important. First, the Mujoco physics engine (Todorov et al., 2012) in which RL agents control a spherical body with continuous steer and roll actions, which sense the position and velocity of nearby others and their own physical state. These agents are rewarded (or penalized) according to their proximity to prey (or predator) agents. Second, we use a 2D gridworld with partial observations (agents locally sense a window of pixels around themselves) as used in Leibo et al. (2017). The former environment allows richer movement patterns while being more challenging to learn in, while the latter allows for very large numbers of agents.

One choice to make in both platforms is the population size at each trophic level. We believe that flocking is more likely when there is a spatial concentration of the predators and/or prey of the species in question. For a species to seek protection from an apex predator we believe it must be hard to escape from its predators because of its density or abilities. Due to these considerations, we opt for a relatively large middle population while a few individual agents at the bottom level represent a large amount of food (e.g. a group of individuals). The apex predators will be the fewest in numbers but have the most impact (on reward).

We find that the agents of the middle trophic level learn a coordinated hunting (flocking) strategy that makes them more successful at hunting their prey. There are at least two benefits from collective hunting; reducing the set of escape trajectories for the prey and collective navigation including information gathering. Cross species collaboration patterns are one of the new possibilities that arise with more than two trophic levels. We observe this in both the 3D physics simulation and the 2D gridworld. The agents at the bottom of the food chain learn to seek out the top apex predator (the hawk) for protection (from jays) and even form a sort of "partnership".

Further, to investigate if at a large scale, like Yang et al. (2018), we also see population dynamics of a form that in some ways resemble nature (e.g. oscillations around a mean), we introduce a variation with spawning and vanishing (from predation) agents. This enables the population levels to reflect the success of the species (policy). We observe several learning phases with lasting equilibria in between quicker changes when superior strategies are discovered. Within each episode, population levels fluctuate regularly around the average, which changes between episodes as all agents learn. Further, we are able to see the aforementioned group behaviours playing a pivotal role. Most interestingly, we see first the failure and then success of group defence without direct individual reward, and it is strengthened by a partnership.

In summary, in our food chain simulations we observe several instances of sophisticated spatial coordination strategies emerging without having shaped the environment dynamics or rewards. For example, we see flocking strategies for predators. While this kind of pattern was also seen recently by Yang et al. (2018) in a grid world, they relied on an explicit ”join group” action and introduce prey explicitly requiring sufficiently large groups to hunt.

Related work

Besides the prior work that has been reviewed above, we here review further relevant literature in Reinforcement Learning, Ecology and Artificial Life.

Predator-Prey dynamics have been widely studied (Levin, 2009), both through data gathering in nature and with mathematical modelling and simulation (Harfoot et al., 2014). Often these models are defined at the population level and deal primarily with numbers or densities in an area. Also, research generally focuses on two trophic levels, a predator species (e.g. foxes) and its prey species (e.g. rabbits). A more intricate line of work (Fretwell, 1987) has considered three or more trophic levels, which permit *trophic cascade effects*. All of the above are explicit mathematical models and do not involve agents that learn.

A famous example of a trophic cascade is the green world hypothesis (Hairston et al., 1960), which explains the richness of plant life on earth as resulting from predation keeping herbivore population size in check. A more recently discovered example (Greeney et al., 2015) is the aforementioned partnership between humming birds and hawks. The natural world contains a tremendous diversity of other intelligent group strategies, including how ants search for food, which has inspired the ant colony optimization class of algorithms (Dorigo, 1992). Flocking has also inspired a long line of work for robot navigation (Reynolds, 1987) enabling drones with weak individual sensors to reach their target more robustly together. Other work has replaced the explicit flocking model with reinforcement learning in an MDP constructed so as to learn flocking (Morihiro et al., 2006), e.g. to fly a group of UAVs in formation to a location (Hung, 2015).

Artificial ecosystems have been studied for a long time (Conrad and Pattee, 1970; Packard, 1987; Ray, 1991; Hraber et al., 1994; Yaeger, 1993; Adami and Brown, 1994). Many of these do not contain an element of spatial navigation. Polyworld developed in Yaeger (1993) is the clearest example that does contain navigation in two dimensions. However, none of these works have reported the emergence of flocking behaviour. Although symbiosis has been a possibility from the earliest models (Conrad and Pattee, 1970), interactions between individuals has not been a main concern (Pachepsky et al., 2002). In their continuation, such as Lenski et al. (2003); Yaeger (2009), these lines of work focused more on the evolution of individual complexity. In the area of artificial life, work on swarm intelligence (Bonabeau et al., 1999) usually involves explicit models as reviewed above. Further, in this area as in this article, work has also focused on generic ecosystems aiming to capture essentials and not biological specifics (Bedau, 2007). Many of these works focus on genetic algorithms due to them being inspired by genetic evolution, while we shift the attention to the sequence of optimization problems addressed by those procedures, and deploy state-of-the-art deep reinforcement learning that have seen much recent success.

Simple predator-prey inspired environments have also been used as test problems for multi-agent reinforcement learning, but not in the same way as explored here. For example, Lowe et al. (2017) model two trophic levels and do not closely investigate the solution strategies, instead comparing algorithms based on accumulated reward.

Reinforcement Learning in Ecosystems

As is common in RL (Sutton and Barto, 1998), we rely on an agent-environment framework (Russell and Norvig, 2010) where an agent interacts sequentially with an environment over a sequence of time steps. The agent selects actions and the environment returns observations and rewards. The agent's performance is measured by cumulative reward, possibly using a discount to encode a preference for the near-term. *Multi-agent reinforcement learning* (MARL) models a collection of agents interacting with an environment and learning, from these interactions, to optimize individual cumulative reward. MARL is typically modelled as a *Markov game* (Littman, 1994). The special case of a Markov game with one agent is a partially-observed Markov decision process (POMDP) (Sutton and Barto, 1998).

The 3D-physics based environment proposed here, features continuous actions, but discrete time. The environment has underlying smooth (classical physics) dynamics with continuous time. 3D physics and continuous control of forces, provide a rich world allowing for more realistic and explicit behaviours, but can be more difficult to learn. For example, it requires a long sequence of actions to perform an apparently simple maneuver.

Ecosystem training We use an ecosystem training (Figure 1) approach where we keep three species (sets of policies), one set each for prey, the predators and apex predator, and for each episode we create a habitat by sampling one policy from each species and use it for all the relevant players. We do this in many parallel threads. Hence, at all times experience is gathered for each policy in many different combinations and for several instances of itself in each ongoing episode. The experiences are gathered and sampled from for each policy, which is learned independently through updates performed to its network weights using state-of-the-art RL algorithms; Maximum a-posteriori Policy Optimization (MPO) (Abdolmaleki et al., 2018) for the continuous case and Impala (Espeholt et al., 2018).

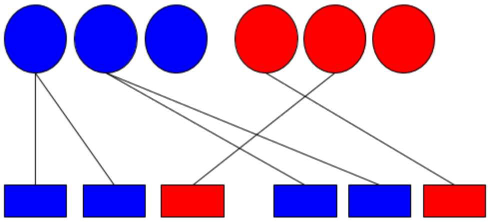

Figure 1: Simple example of ecosystem training: Sampling from two species (blue and red) of three policies (here circles) each, to be placed in habitats (here row of three rectangles) for two identical blue and one red.

A Physics-Based Food Chain Environment

We introduce a physics based food chain environment (see Figures 2 and 4) and describe its observations and rewards, as well as how agents process these and learn in an ecosystem training framework with three species. We base our environment on the MuJoCo physics engine (Todorov et al., 2012), utilized in much recent continuous reinforcement learning work including Brockman et al. (2016); Heess et al. (2017); Bansal et al. (2017); Abdolmaleki et al. (2018). In this environment, each agent controls a sphere with a two dimensional action space; acceleration forward/backward and rotational to steer. We have three different roles in the environment; apex predator, predator and prey, so we have three agent types or species. For visualization we render each agent type with different colors; green(prey), blue(predator) and red(apex predator). These spheres travel on a square floor bounded by walls on each side. The environment further contains two large square blocks which serve as physical barriers and introduce additional structure in the environment. The predators always spawn randomly within a large square in the middle. The apex predator and prey spawns according to two equally likely patterns. In the first they spawn in the same central square as the predator, or they each spawn independently in (a square in) a uniformly random corner. Both spawn patterns are displayed in accompanying videos[1] and are simply chosen to force the learning of varied behaviours, but are not designed to generate any specific outcome. In one, the predators (and the apex predator) has to start with searching, in the other the prey has to start with escaping.

Proximity based rewards The agents receive rewards based on proximity to other agents. Predators receive positive reward for being near prey while prey agents receive negative reward. Similarly, the apex predator receives a positive reward when it is sufficiently close to a predator agent which receives a negative reward in turn. The reward function is only dependent on distance between the agents and

[1] https://docs.google.com/presentation/d/1u86oapziZ35MfphcrIC3zbMMg9It-Bhf6fJEqyoeBwg/edit?usp=sharing

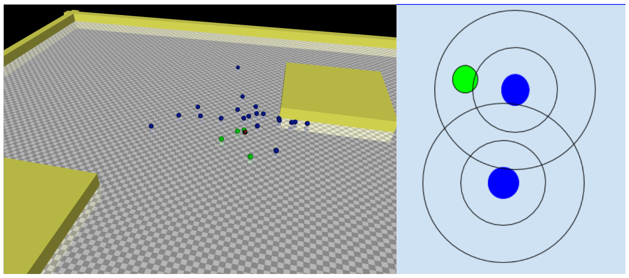

Figure 2: Physics based ecosystem environment. Left: An example of close red-green partnership in the middle of a large floor. Right: An illustration of two predators (blue) with a larger conspecific radius and a smaller by which it can see the prey (green).

we choose a sigmoid with a cut-off threshold. We define,

$$\phi(d) = \begin{cases} 1 - \tanh(0.5d) \; \textit{if} \; d < radius \\ 0 \; \textit{otherwise} \end{cases} \tag{1}$$

and if d_i is the distance (at time t) from a certain predator agent to the prey agent i, and if e_j is the distance to the apex predator agent j, then this predator agent receives the instantaneous reward $\sum_i \phi(d_i) - 2.5 \sum_j \phi(e_j)$. The factor 2.5 for the term that represents being predated on by the apex predator, is to not make being close to one or two prey more important then avoiding the predator. A prey agent's reward only depends on its distance to the predator agents, and if the distance to the predator agent l is d_l, then its instantaneous reward is $-\sum_l \phi(d_l)$. An apex predators agent's reward only depends on its distance to the predator agents, and if the distance to the predator agent l is e_l, then its instantaneous reward is $\sum_l \phi(e_l)$.

Observations Every agent observes their own position, velocity and accelerometer information as well as the vector to each corner of the two blocks that can be seen in the figure. Each agent further observes egocentrically represented positions and velocities of other agents within its sensor radius. Next, we introduce how the agents process these observations and map onto actions.

Perception Network As is common with swarm agents, motivated by both biological inspiration as well as learning complexity, we only want our agents to take the species of another agent into account and not individual identity. Further, we let each agent in the environment of the same species have the same policy. We want agents to have the capacity to generalize across different numbers of sensed other agents and potentially scale to very large numbers. We achieve this, as shown in Figure 3, by first applying a two layer feed forward neural network to every other agent's position and velocity (within the focal agent's sensor radius), and then we can combine these for each agent type by either summing or computing the mean, with similarity to Hüttenrauch et al. (2017).

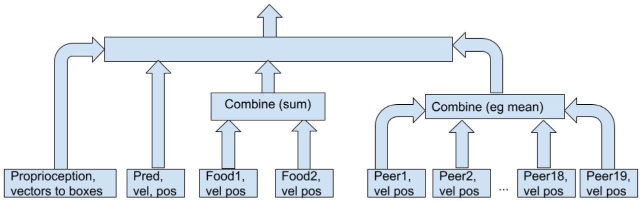

Figure 3: Diagram showing how the perception network first process the different parts of an observation separately and then combines the resulting representations.

Policy Network The policy that produces the continuous action vector is constructed as follows. The network that encodes the position and velocity of each other agent within the relevant sensor radius, is a two-layer feed forward network with 8 hidden units in each layer. For the predator agents (as many as 40 in the physics based experiments), we combine these representations by computing their mean. For the prey and apex predator agents, which are few (5 and 1), we use the sum to distinguish different numbers of agents in the same place. After this, we concatenate the result for the three types of other agents as well as the agent's own proprioception and the vectors to the corners of the boxes on the floor. The resulting total representation is first processed by a two layer feedforward network with 128 and 64 units (with tanh activation) and then a recurrent network, an LSTM (Hochreiter and Schmidhuber, 1997) with 32 units. The final layer produces Gaussian distributed actions. Note that while parameters are shared between agents of the same type within an episode, agents are entirely independent in terms of action selection.

Locally observed grid worlds

In this section, we introduce a grid world ecosystem that is similar to the 3D physics based world, in the abstract ecological sense. The agents environment is a square map and they can rotate 90 degrees left or right, step forward or backward, or launch a yellow beam that represents predation. The predation beam is a difference to the physical 3D simulation where an agent only has to be near its prey to predate. In the grid world the agent has to be near and directed towards the prey, and choose this action, which is supplied by the platform Leibo et al. (2017). It is our strategy to only make as minimal and obvious design choices on top of the generic platforms as possible. However, it also comes with new possibilities including the possibility of enabling defense against predation and it makes a form of capture of prey even more important. Again, we have agents of three different varieties; prey(green), predator(blue) and apex predator(red). The apex predator gets reward $+1$ if predating on a predator (meaning that the predator is in the apex predator's yellow beam), while the predator gets -1. Sim-

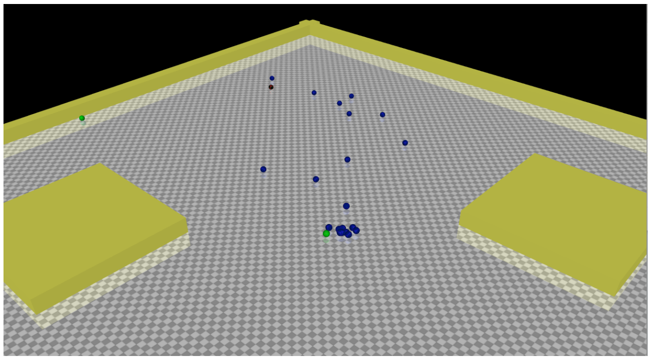

Figure 4: One prey (green) getting surrounded by most of the predators (blue).

ilarly a predator gets $+1$ for predating on prey, which then gets -1. These are the rules that defines the environment shown in Figure 5.

The agents are trained using a ecosystem training protocol like in the 3D physics case, while differing by the learning update used. We replace MPO with impala learning updates (Espeholt et al., 2018) for this case in which the action space is discrete. The agents observe a small 9x9 window around the agent. Figure 5 shows screens from the resulting simulations when using a modest 5 apex predators, 10 prey and 50 predators. In the next section, we will also consider agents that spawn and vanish during very long episodes with thousands of agents, yielding population dynamics reflecting the relative success of each species over time.

Experiments

This section presents results from experiments with the introduced ecosystems that test our hypothesis that flocking and symbiosis can result in response to very generic predator-prey pressures in a spatial environment with range-limited sensation and action. We also investigate how the emergence of the relevant strategies depends on levels of predation pressure and the range of conspecific (within species) sensing among the predators.

Physical worlds

Our first range of experiments in ecosystems with 3D-physics, is varying the radius (5 vs 10 in a square with side length 48) within which predator agents can see each other, to see how well they make use of that information and what the consequences for the ecosystem are. All experiments features a radius of 5 both for sensing agents of other types, for the apex predator and the prey to see their conspecifics and the environment reward cut-off radius. We measure both the reward achieved for each type of agent during training, which was performed with 200 parallel habitats, and during regular evaluation against fixed pretrained prey and apex predator agents.

conspecific radius	5	10
number of predators	20	20
evaluation reward	2.7 ± 0.1	5.1 ± 0.2
symbiosis	0.6 ± 0.2	1.5 ± 0.4
predator group size	5.5 ± 1.4	9.3 ± 2.5
apex predator rewards	29.7 ± 5.8	51.8 ± 5.1
predator rewards	5.6 ± 1.5	9.3 ± 0.4
prey rewards	-36.0 ± 4.6	-64.0 ± 13

conspecific radius	10	10
number of predators	40	10
symbiosis	1.7 ± 0.3	0.96 ± 0.22
predator group size	16.8 ± 2.8	4.2 ± 1.3
apex predator rewards	119 ± 20	25.9 ± 8.7
predator rewards	12.9 ± 8.7	8.5 ± 5.2
prey rewards	-128 ± 18	-22.2 ± 9.1

Table 1: Results in 3D-physics ecosystems (at the end of training) for evaluation predator reward against fixed apex predator and prey, as well as on symbiosis measured by the average number of prey near the apex predator, predator group size measured by the average number of other predators near a predator and training reward for each species. The result are from three full runs of each of four experiments, pooling results for each species (3x10 policies) and calculating averages and standard errors. For reward, it is just the standard error for the three means. The ecosystems features 1 apex predator, 10, 20 or 40 predators, 5 prey and conspecific radius 5 or 10.

Unlike the results against the prey and the apex predator that the predators are learning with and that keeps changing, the results against the pre-trained agents provide a consistent well-defined evaluation. Further, during training we also measure distances between pairs of predator agents as well as between the apex predator and prey pairs of agents. From these we can judge to what extent predator agents flock and to what extent prey and the apex predator stay together. The results can be found in Table 1 based on species averages over the last $0.5e10$ of $3.0e10$ training steps.

Predator coordination: Predators perform better when able to sense other predators with a larger sensor radius as can be seen in Table 1 ($5.1 > 2.7$). We also see larger predator groups ($9.3 > 5.3$) when the conspecific sensor radius is larger. We believe that where there is a larger concentration of predators, there is likely to be a prey, and it is easier to join such groups if one can sense further.

Apex predator-prey coordination: Another observation is that the number of prey on average within a circle around the apex predator is much higher ($1.5 > 0.6$) when the predator has the wider conspecific sensor radius, which makes them a more effective hunter and increases the incentive of protection for prey. The smaller number (0.6) can

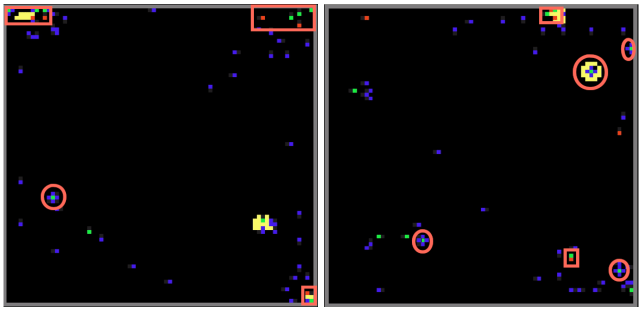

Figure 5: Grid world ecosystem showing group captures (predators of prey, highlighted by orange ellipses) and the apex predator-prey partnerships (orange rectangles).

largely be due to the apex predator chasing a group of predators that is chasing a prey and thereby, keeping the prey and the apex predator near each other. The stronger pattern (1.5) is clearly a more direct partnership pattern as can be seen in the videos[1], from which Figure 2 was taken. When the prey is chased by a flock of predators it cannot escape, it heads to the middle where it meets the apex predator that breaks up the pursuing predators. The videos[1] also show an example with a smaller world where the two prey and the apex predator find each other and the three stay together.

It is interesting that the apex predator earns substantially more reward when this partnership pattern with prey has emerged. The end result of predators getting better at hunting is that its own predator (the apex predator) is a big winner, via the adaptation by the prey. This is an example of the fascinating and indirect possibilities that arise from modeling more than two trophic levels. Figure 2 shows an example situation at the end of the learning for this case, while Figure 4 shows a situation generated by the very same agents (same weights) as in Figure 2, but here a lone prey is surrounded by a very large number of predators and does not find an escape route until the apex predator agent arrives. The prey here suffers catastrophic reward.

Varying numbers of predators: We compare varying numbers of predators, all with conspecific sensing radius 10. We see (Table 1) that the apex predator-prey partnership emerges more with 40 predators when the pressure on prey is obviously higher, and much less with 10 predators.

Grid worlds

In a first grid world ecosystem experiment with 5 apex predators, 10 prey and 50 predators, we consistently see groups of predators capturing prey agents, both in the middle of the floor (by 4 predators) and against walls (3 predators) or in corners (2 predators). We also see prey learning the defence strategy of sheltering near the apex predator, and the apex predator staying with the prey as it can enjoy reward for predating on approaching predators. These two strategies are visible multiple times in Figure 5.

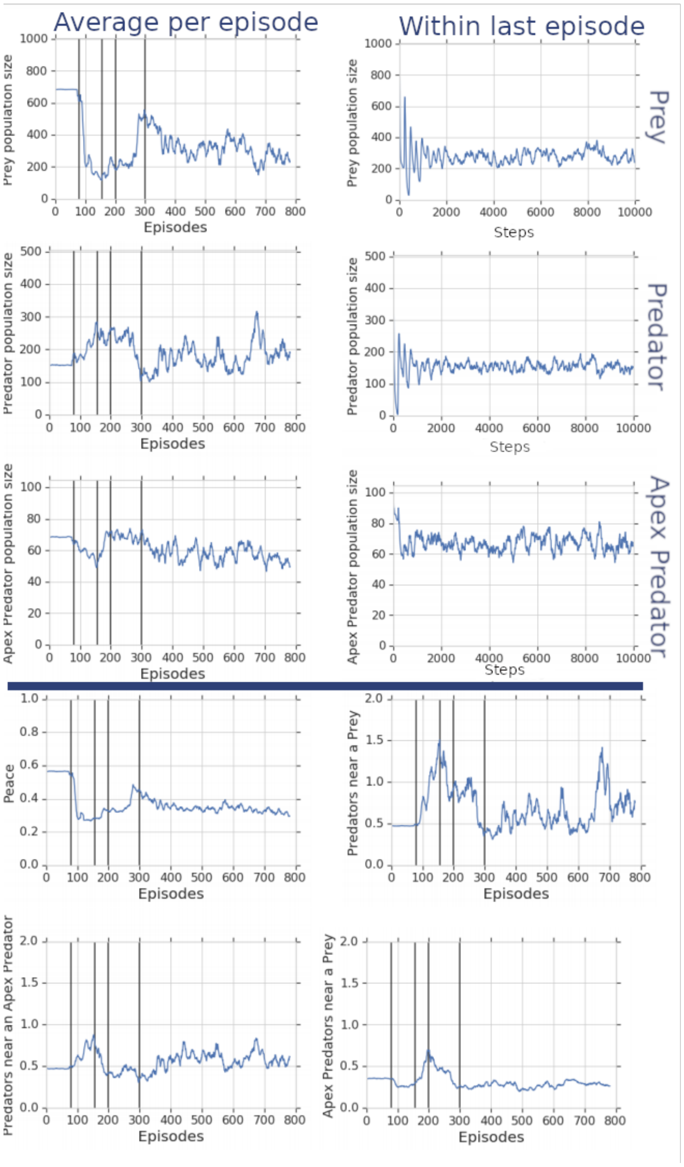

Figure 6: Top three on the left:The number of prey (top), predators (middle) and apex predators (bottom) present on average per episode. Top three on the right: The number of prey (top), predators (middle) and the apex predator (bottom) during the last episode. Bottom two rows: Peace (fraction of time agents on average is present in the environment (not vanished), numbers of predators near (within 3 cells on each side) prey, apex predators near prey (sheltering) and predators near the apex predator.

Large scale worlds with spawning and vanishing agents. In our final experiment, we extend our experiments both in scale and to introduce population dynamics, in the simplest way available. We give each agent a total amount of health (5) at the start, which is depleted by one unit each time they are preyed upon. Upon depletion, agents sit out of the game for a certain amount of time (200) and then are respawned. While neither this, nor a fixed spawn rate as in Yang et al. (2018), represents the multiplicative nature of biological population dynamics, we do still get an environment where the relative population size indicates how well a certain agent type performs at a point in learning history. Within episodes we get a fluctuating curve with a significant oscillatory component around this mean, as is also familiar from ecology (Levin, 2009) if not in an identical manner.

In addition, our design allows for defence by prey species. That is, a predator can fire its yellow beam at the apex predator to reduce its health by one, and similarly prey can defend against predators. The structure of this defence indirectly encourages group strategies, since it takes a while for defence to eliminate the predator. While better for all agents if predators are eliminated, each individual can selfishly optimize reward by running away and allowing one's peers to defend them. While reminiscent of the sequential social dilemmas of Leibo et al. (2017), here a solution involving a familiar partnership with the third species is found.

From the perspective of the prey, we see a progression from annihilation to initiating defence to seeking shelter near the apex predator and jointly decimating the predator, which decrease its predation and peace increase. After this, the apex predator must hunt the predator more actively, and the predator gradually increase its predation on the prey.

Figure 6 showz the result of this experiment. Each (very long 10000 step) episode starts with 1000 prey, 500 predators and 100 apex predators, in terms of population level at different times of learning, and how the numbers varies within an episode. After the first 100 (parallel) episodes (24 hours on the compute cluster), we see a drop in prey numbers to near extinction, and in the within episode results, we see a complete annihilation early in the episodes. Predators have learnt early on to efficiently hunt prey. Closer inspection shows that the prey mostly individually flee, while they also start using their beam at the predator. After a period of nearly 300 episodes (3 days on cluster), prey numbers quickly increase while predator numbers now plummet before they adapt to this situation and gradually improve at the expense of both prey and apex predators. In Figure 6 (bottom right) we can see that the number of apex predators on average around a prey, climbs up to a peak just at the time when the change starts around episode 200. After this, we can see that this is followed by peace (low predation) to increase. We see predation levels between all species decrease substantially and we also see in Figure 6 that the species are now less frequently near each other. Our interpretation is that the predators decreased their pursuit of prey as a response to an apex predator-prey partnership that made it face a combination of predation and defence it was decimated by.

Conclusions

In this work, we have approached emergence of complex group behaviours in ecosystems as a sequence of optimization problems where each species is optimizing its fitness based on the current policies of other species. We use state-of-the-art deep reinforcement learning methods to address these optimization problems and consistently found the hypothesized emergence of strategies like flocking and symbiosis. These group patterns, which appear widely in different forms and contexts in nature, emerge through interactions of independently incentivized self-interested reinforcement learning agents acting in simple ecosystems that are not shaped to encourage the emergence of these particular strategies. Taken together, our results demonstrate that state-of-the-art reinforcement learning agents combined with our open-ended ecosystem training protocol can generate interesting coordinated behaviours familiar from nature.

References

Abdolmaleki, A., Springenberg, J. T., Tassa, Y., Munos, R., Heess, N., and Riedmiller, M. (2018). Maximum a posteriori policy optimisation. In *ICLR 2018*.

Adami, C. and Brown, T. C. (1994). Evolutionary Learning in the 2D Artificial Life System Avida. In Brooks, R. A. and Maes, P., editors, *Artificial Life IV*, pages 377–381, Cambridge, MA. MIT Press.

Alvard, M. S. (2003). Kinship, lineage, and an evolutionary perspective on cooperative hunting groups in Indonesia. *Human Nature*, 14 2:129–63.

Ashby, W. (1947). Principles of the self-organizing dynamic system. *The J. of General Psychology*, 37(2):125–128.

Bansal, T., Pachocki, J., Sidor, S., Sutskever, I., and Mordatch, I. (2017). Emergent complexity via multi-agent competition. *arXiv preprint arXiv:1710.03748*.

Bedau, M. A. (2007). Artificial life. In Matthen, M. and Stephens, C., editors, *Philosophy of Biology*, Handbook of the Philosophy of Science, pages 585 – 603. North-Holland, Amsterdam.

Bonabeau, E., de Recherches Du Fnrs Marco Dorigo, D., Dorigo, M., Théraulaz, G., and Theraulaz, G. (1999). *Swarm Intelligence: From Natural to Artificial Systems*. Proceedings volume in the Santa Fe Institute Studies in the Sciences of Complexity. OUP USA.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). OpenAI Gym. *CoRR*, abs/1606.01540.

Conrad, M. and Pattee, H. (1970). Evolution experiments with an artificial ecosystem. *Journal of Theoretical Biology*, 28(3):393 – 409.

Dorigo, M. (1992). *Optimization, Learning and Natural Algorithms*. PhD thesis, Politecnico di Milano, Italy.

Espeholt, L., Soyer, H., and Munos, R. e. a. (2018). IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1407–1416.

Fretwell, S. D. (1987). Food chain dynamics: The central theory of ecology? *Oikos*, 50 3:291–301.

Greeney, H. F., Meneses, M. R., Hamilton, C. E., Lichter-Marck, E., Mannan, R. W., Snyder, N., Snyder, H., Wethington, S. M., and Dyer, L. A. (2015). Trait-mediated trophic cascade creates enemy-free space for nesting hummingbirds. *Science Advances*, 1(8).

Hairston, N. G., Smith, F. E., and Slobodkin, L. B. (1960). Community structure, population control, and competition. *The American Naturalist*, 94(879):421–425.

Handegard, N. O., Boswell, K. M., Ioannou, C. C., Leblanc, S., Tj, D., and Couzin, I. D. (2012). The dynamics of coordinated group hunting and collective information transfer among schooling prey. *Current Biology*, 22:1213–1217.

Harfoot, M. B. J., Newbold, T., and Tittensor, D. P. e. a. (2014). Emergent Global Patterns of Ecosystem Structure and Function from a Mechanistic General Ecosystem Model. *PLoS Biol*, 12(4):e1001841+.

Heess, N., TB, D., Sriram, S., Lemmon, J., Merel, J., Wayne, G., Tassa, Y., Erez, T., Wang, Z., Eslami, S. M. A., Riedmiller, M. A., and Silver, D. (2017). Emergence of locomotion behaviours in rich environments. *CoRR*, abs/1707.02286.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Hraber, P. T., Jones, T., Forrest, S., and Ds, K. (1994). The ecology of echo. *Artificial Life*, pages 165–190.

Hung, D. S. M. (2015). *Reinforcement Learning Approach to Flocking with Fixed-Wing UAVS in a Stochastic Environments*. PhD thesis, Royal Military College of Canada.

Hüttenrauch, M., Sosic, A., and Neumann, G. (2017). Guided deep reinforcement learning for swarm systems. *CoRR*, abs/1709.06011.

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 International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017)*.

Lenski, R. E., Ofria, C., Pennock, R. T., and Adami, C. (2003). The evolutionary origin of complex features. *Nature*, 423:139–144.

Levin, S. A. (2009). *The Princeton Guide to Ecology.* Princeton University Press.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *In Proceedings of the Eleventh International Conference on Machine Learning*, pages 157–163. Morgan Kaufmann.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., and Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. *CoRR*, abs/1706.02275.

Morihiro, K., Isokawa, T., Nishimura, H., and Matsui, N. (2006). Emergence of flocking behavior based on reinforcement learning. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 699–706, Berlin, Heidelberg. Springer Berlin Heidelberg.

Pachepsky, E., Taylor, T., and Jones, S. (2002). Mutualism promotes diversity and stability in a simple artificial ecosystem. *Artificial life*, 8:5–24.

Packard, N. H. (1987). Evolving bugs in a simulated ecosystem. In Langton, C., editor, *Proceedings of the Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems (ALIFE '87)*, Santa Fe Institute Studies in the Sciences of Complexity, pages 141–156.

Ray, T. (1991). An approach to the synthesis of life. In *Artificial Life II*, pages 371–408. Addison-Wesley.

Reynolds, C. W. (1987). Flocks, herds, and schools: A distributed behavioral model. *SIGGRAPH Computer Graphics*, 21(4):25–34.

Russell, S. J. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach.* Prentice Hall, Englewood Cliffs, NJ, 3^{rd} edition.

Ruxton, G. (2012). Group dynamics: Predators and prey get a little help from their friends. *Current Biology*, 22(13):R531 – R533.

Sutton, R. and Barto, A. (1998). *Reinforcement Learning: An Introduction.* MIT Press.

Todorov, E., Erez, T., and Tassa, Y. (2012). Mujoco: A physics engine for model-based control. *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033.

Yaeger, L. (1993). Computational genetics, physiology, metabolism, neural systems, learning, vision, and behavior or polyworld: Life in a new context. In *Artificial Life III, Vol. XVII of SFI Studies in the Sciences of Complexity, Santa Fe Institute*, pages 263–298. Addison-Wesley.

Yaeger, L. S. (2009). How evolution guides complexity. *HFSP Journal*, 3(5):328–339.

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 '18.