Emergent Heterogeneous Strategies from Homogeneous Capabilities in Multi-Agent Systems

Rolando Fernandez[1], Erin Zaroukian[1] James D. Humann[1], Brandon Perelman[1], Michael R. Dorothy[1], Sebastian S. Rodriguez[2], and Derrik E. Asher[1]

[1] US CCDC Army Research Laboratory, Adelphi MD 20783, USA,
rolando.fernandez1.civ@mail.mil,
[2] Department of Computer Science, University of Illinois at Urbana-Champaign,
Urbana IL 61801, USA

Abstract. In multi-agent systems, agents' abilities are often used to classify a system as either homogeneous or heterogeneous. In the context of multi-agent reinforcement learning (MARL) systems, the agents can also be homogeneous or heterogeneous in their strategies. In this work, we explore instances where agents with homogeneous capabilities must collaborate to achieve a common goal in a predator-prey pursuit task. We show that results from homogeneous and heterogeneous strategies associated with learning differ substantially from agents with fixed strategies that are analytically defined. Given that our agents are homogeneous in capability, here we explore the impact of homogeneous and heterogeneous strategies in a MARL paradigm.

Keywords: homogeneous, heterogeneous, strategy, predator-prey pursuit, multi-agent reinforcement learning, simulation

1 Introduction

1.1 Homogeneity and Heterogeneity in Teams

Degree of heterogeneity is a major differentiating factor among multi-agent systems [16]. Like human organizations, agent teams are often formed to take advantage of supplementary similarities or complementary differences [7]. Similarity is leveraged by scaling the system size up with homogeneous agents that can work in parallel to increase the task completion rate. Differentiation in heterogeneous systems allows for specialization to complete diverse sub-tasks that can be integrated into completion of the full task. For example, in the case of human-AI centaur chess teams, amateur human players and their AI teammates are able to achieve better performance than either human grandmasters or supercomputers, by crafting their method of interaction to leverage one another's strengths [6].

1.2 Sources of Heterogeneity

In addition to heterogeneity based on form factor (e.g., a team of ground and aerial robots) or hardware-defined function [14], there are agent teams with identical hardware but heterogeneous behaviors. For example, it has been shown that a team of 5 robots with identical hardware, whose labor was divided between digging (prying boxes away from the wall) and twisting (clustering boxes in the center of the testbed), was able to cluster groups of boxes more efficiently than homogeneously programmed agent teams [15]. Heterogeneous behavior can also be achieved by dynamic state-switching, where agents assume different roles based on their local perception of the environment and task needs, even if they are all running the same behavioral algorithms [8]. Heterogeneity has also been shown to naturally emerge from homogeneous capabilities: in some insect species, juvenile members can be differentially nurtured so that they show marked dimorphism at maturity, enabling differentiation into soldier ants and drones [11]. So we see that heterogeneity can substantially improve system performance and can emerge from various sources, such as nurture or training, hardware, algorithms, or local time-dependent behavior differentiation.

1.3 Heterogeneity through Reinforcement Learning

The source of heterogeneity we study here is trained behavioral heterogeneity from a reinforcement learning (RL) approach. Multi-agent systems may be trained to develop heterogeneous, individual, algorithms for each agent; train as a set with mutually known inputs and actions; or train homogeneously but with differing sensor information so that agents make decisions locally.

This raises interesting questions. If multiple agents are trained to complete a task over successive trials, do they learn to differentiate their behavior and take advantage of complementary differences? If so, does each agent learn a fixed role, or are the roles distributed dynamically? If a teammate is lost, changed, or compromised, have the other agents learned robust strategies to compensate?

1.4 Collaboration in Multi-agent Systems

Changing the reward structure in RL from zero-sum to shared rewards can cause qualitatively different behaviors to emerge from the learning agents [17], implying that they are learning to cooperate. Even when agents are allowed to train heterogeneously, it is difficult to definitively say that they are learning to specialize or even consider their teammates [1]. Instead of collaborating, they may simply be learning to maximize their own reward in a way that generally scales well to group settings (i.e. learning complementary similarities even if supplementary differences are possible) or find strategies that perform well irrespective of their teammates' actions.

Collaboration between cooperative agents in multi-agent systems is often poorly defined. For our purposes, we define collaboration with two fundamental components: 1) measurable coordination between agents actions (e.g., a measurable quantity that indicates agents actions have inter-dependencies), and 2) inferred cooperation that can be determined by an alignment of agents' efforts or a common goal. A clear definition of collaboration allows us to objectively determine the emergence of this phenomenon in addition to measuring its strength. Further, typical analysis of collaboration is restricted to task-performance, which does not explicitly allow one to quantify differences in collaborative efforts (e.g., if two groups perform the same, can we conclude that they are both collaborating in the same way?). Finally, explicit quantification of collaboration may permit predictions of group strategy or group behaviors when novel partners are introduced (e.g., the inclusion of human partners into a multi-agent system). Therefore, we have defined collaboration in such a way to quantify differences between homo- and heterogeneous strategies in multi-agent systems.

1.5 Overview of Paper

In the following sections, the concept of heterogeneous strategies emerging from homogeneous capabilities is demonstrated. In the Methods section, we describe our simulation environment, where we test agents that are guided either by a learning algorithm or fixed strategies. Next, agent performance is shown with probability distributions and statistics for both homogeneous and heterogeneous cases in the Results section. Finally, the Discussion section points to the conclusions that were drawn from the results and provides further avenues of research associated with heterogeneous strategies from homogeneous capabilities in multi-agent systems.

2 Methods

2.1 Simulation Environment

A continuous bounded 2D simulation environment was utilized to train and evaluate a set of four agents (three predators and one prey, represented as circles) per model in the predator-prey pursuit task [4, 10] and a visualization of the task is shown in Fig. 1. The predators scored points (i.e., were given reward during training and evaluated during testing) every time they collided with the prey. Predator agents were homogeneous in their capabilities (i.e., same size, velocity, and acceleration limitations), whereas the prey was 33% smaller, could accelerate 33% faster, and had a 30% max speed advantage. The simulation environment was built upon the OpenAI Gym library [5] and developed for use with the multi-agent deep reinforcement learning algorithm, multi-agent deep deterministic policy gradient (MADDPG) [10]. We assume that the predator agents must cooperate to score a hit on the prey, given the prey's capability advantages. Prior work has shown that a simple greedy policy (i.e., minimize distance to prey) is insufficient for the predators to succeed [3].

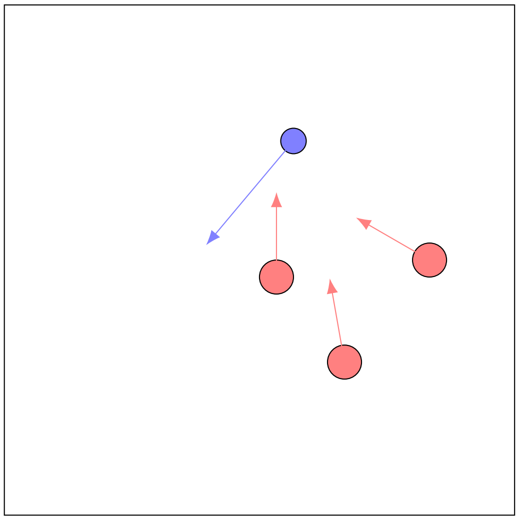


Fig. 1: Predator-prey pursuit particle environment

All the agents were trained concurrently using the MADDPG algorithm [10]. MADDPG utilizes a decentralized-actor centralized-critic framework which accounts for each agent's observations and actions during training. The predators all received the same fixed reward (shared/joint reward) when any one of them hit the prey, while the prey received the negative of the same fixed reward when hit. At the start of each episode the initial positions of the agents were randomized, and their initial accelerations and velocities were set to zero. The state space of each predator agent contained its absolute velocity, absolute position in the environment, relative distance and direction to the other predators and the prey, and the prey's absolute velocity. The state space of the prey agent contained its velocity, absolute position in the environment, and relative distance to the predators. The action outputs of the policy network for an agent are accelerations in the two-dimensional coordinate system.

2.2 Analytically-defined Agents

In addition to the MADDPG-trained agents, we consider two analytically-defined agents (also referred to in this article as fixed-strategy agents), called a Chaser agent and an Interceptor agent. The Chaser agent does not leverage any prey velocity information and only points its own velocity directly at the prey's instantaneous position. In contrast, the Interceptor agent considers both instantaneous position and velocity of the prey. At a moment in time, the Apollonius circle describes the potential interception locations if both agents continue in a constant direction [9]. Given a prey that is faster than the predator, only a subset of possible constant prey strategies admit a capture trajectory for the predator [12]. We extend the Apollonius circle strategy for the predator in the case where capture is not possible. The case where capture is possible is shown in Fig. 2a. Equal travel time at capture gives $\frac{d_E}{V_E} = \frac{d_P}{V_P}$, and the rule of sines gives $\frac{d_P}{\sin\phi} = \frac{d_E}{\sin\theta}$, resulting in

$$\sin\theta = \frac{V_E}{V_P}\sin\phi. \tag{1}$$

In the case where capture is not possible, we consider a finite time prediction for prey trajectory and choose the predator's strategy to minimize the final distance. This is shown in Fig. 2b, where the d_P circle represents how far the pursuer is able to travel in the time it took the evader to travel distance R. The optimum position for the purser to minimize the relative distance at the moment the evader reaches the R circle is to head straight toward that point. Clearly, $\psi = \pi - \theta - \phi$, and

$$\frac{\sin\theta}{R} = \frac{\sin(\theta+\phi)}{r}. \tag{2}$$

The critical case between the capture set and non-capture set is $\sin\phi^* = \frac{V_P}{V_E}$, $\theta^* = \frac{\pi}{2}$, and evaluating Eq. 2 at that point gives

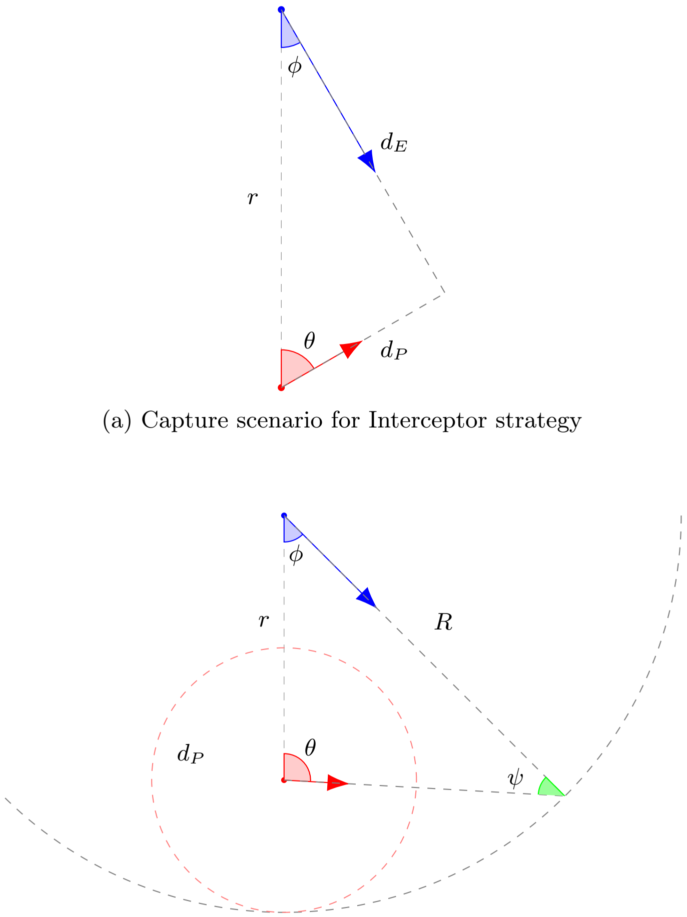
$$R = \frac{r}{\sqrt{1-\frac{V_P^2}{V_E^2}}}. \tag{3}$$

This value for R will result in a continuous policy across all cases.

2.3 Experimental Design

To compare heterogeneous and homogeneous team structures, the RL agents and fixed-strategy agents were subdivided into homogeneous and heterogeneous teams. For the RL agents, the homogeneous teams consisted of a single predator agent's policy (or strategy) duplicated and implemented onto the other two predators (i.e., 3 separate homogeneous teams per model). Whereas, the RL agent heterogeneous teams consisted of the original policies or strategies that emerged through training. Homogeneous and heterogeneous teams for fixed-strategy predator agents are intuitively labeled.

Two previously trained models, independently trained with the MADDPG algorithm for 100,000 episodes and 25 timesteps per episode [2], and used them for our evaluations. Trained Model 1 and 2 predator agents were tested as heterogeneous teams consisting of all three agents (i.e., Agents 0, 1, and 2), labeled as 'All Agents' in Figures 3a and 3b, or as homogeneous teams in which all three actors' behaviors are driven by the policies of either Agent 0, Agent 1, or Agent 2 from each of the models. Similarly, the fixed-strategy predator agents were tested against both Model 1 and Model 2 prey agents in several types of homogeneous or heterogeneous team compositions: 3 Interceptors, 2 Interceptors and 1 Chaser, 1 Interceptor and 2 Chasers, and 3 Chasers. All tests were performed for 1000 episodes at 1000 timesteps per episode.



(a) Capture scenario for Interceptor strategy

(b) Finite R Interceptor strategy

Fig. 2: Pictorial of Interceptor strategy for capture and non-capture cases

3 Results

3.1 Learned Policy Agent Experiments

The simulation results show how group performance changes upon replicating a single agent's policy network and thus introducing homogeneous strategies. Further, the performance resulting from this homogeneous strategy implementation may provide a means of classifying different trained policies that emerge through collaboration in the multi-agent reinforcement learning process.

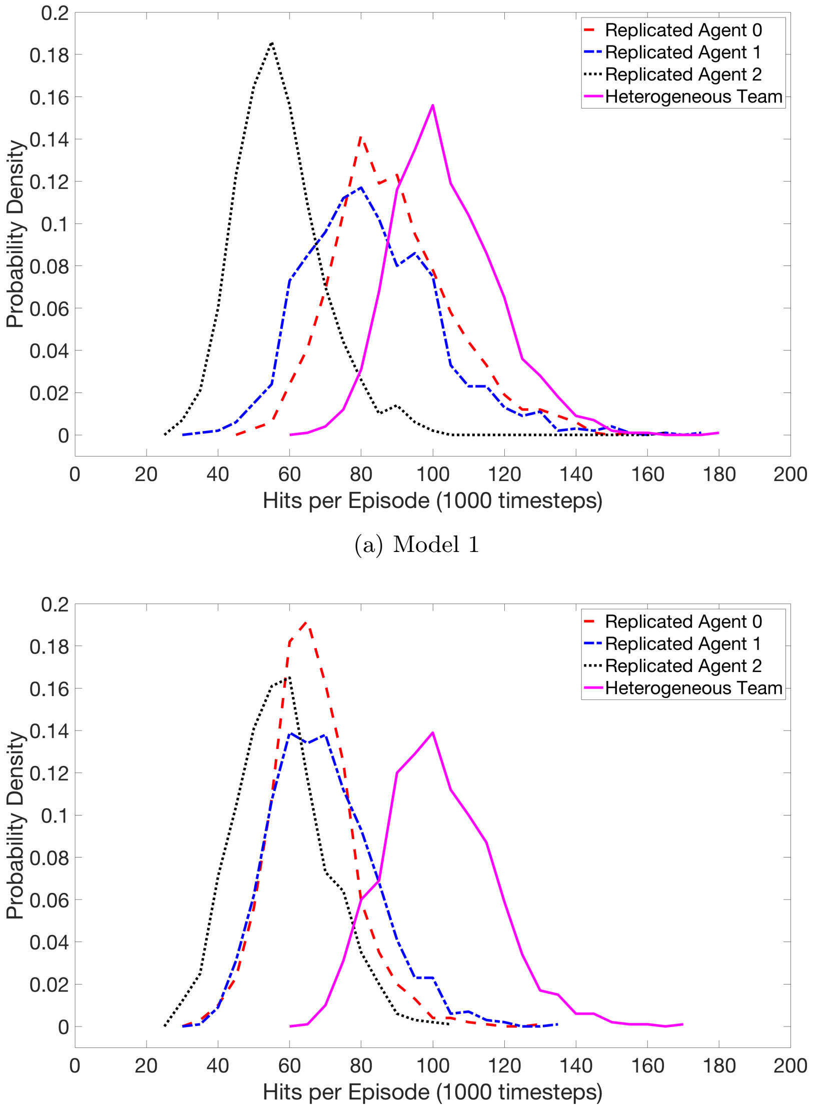
Using the data collected during testing we generated probability density plots of the hits the predators achieved on the prey to analyze agent performance for each of the models (Figures 3a and 3b). 'All Agents' shows the aggregated or team performance of three predator agents with their independently trained network policies. 'Agent 0' represents the data generated from the replication of 'Agent 0' across the three predator agents. Similarly, 'Agent 1' and 'Agent 2' respectively represent the replication of their corresponding agent policies. The x-axes show the number of hits (i.e., number of times the predators collaboratively contacted the prey agent throughout an episode). The y-axes show the normalized frequency or probability density for the hits per episode. We performed the pairwise 2-sample Kolmogorov-Smirnov (KS) test to show that all the distributions were significantly different from one another at the alpha = 0.01 level (p-values « 0.001).

We can see from Figures 3a and 3b, that in the case of the MADDPG-trained agent strategies the heterogeneous predator team (All Agents) is able to outperform all of the homogeneous predator teams (Agent 0, Agent 1, and Agent 2) where the same policy is replicated across each agent. Furthermore, as all the homogeneous team performances were significantly different we can infer that each individual agent policy learned to utilize a different strategy that benefited the team as whole.

3.2 Analytically-defined Agent Experiments

The data in Tables 1a and 1b shows the performance statistics for the fixed-strategy predator agents playing against the MADDPG-trained (RL) prey. The statistics were generated by fitting an exponential distribution to the probability density functions of the data from the respective cases. Note, these data are shown in tables rather than plots as the probability density functions are all visually similar when presented with x-axis values on the same scale as Figures 3a and 3b.

Overall, the tables allow us to see that the performance of the Chaser predators is an order of magnitude worse than that of the Interceptor predators when playing against trained prey from Models 1 and 2, which was also shown to be significantly different with 2-sample KS-tests ($p < 0.001$). Interestingly, when we combine a Chaser predator with two Interceptor predators (i.e., heterogeneous fixed-strategy cases), the inclusion of Chaser predators results in a significant reduction in group performance (compare 3 Interceptors case to the heterogeneous



(a) Model 1

(b) Model 2

Fig. 3: Probability density performance plots for learning predators

cases in Table 1). In addition, across both models, combining an Interceptor predator with two Chaser predators significantly improves group performance from the three Chasers case ($p < 0.001$). Together, these results suggest that overall the Interceptor strategy is significantly better than the Chaser strategy, especially in the homogeneous cases.

Table 1: Performance of fixed strategy predators

Predators' Strategy	**Mean**	**CI Lower**	**CI Upper**
3 Interceptors	1.55	1.46	1.65
2 Interceptors, 1 Chaser	1.16	1.09	1.24
1 Interceptors, 2 Chaser	0.876	0.824	0.932
3 Chasers	0.154	0.145	0.164

(a) Model 1

Predators' Strategy	**Mean**	**CI Lower**	**CI Upper**
3 Interceptors	1.44	1.35	1.53
2 Interceptors, 1 Chaser	1.05	0.990	1.12
1 Interceptors, 2 Chaser	0.998	0.939	1.06
3 Chasers	0.144	0.136	0.153

(b) Model 2

4 Discussion

Above we compared performance in a predator-prey pursuit task for agents using either homogeneous or heterogeneous policies or strategies.[3] The heterogeneous strategies were either created through independently-trained learning agents or by assigning different analytical strategies (Chaser and Interceptor) to different predator teammates. The homogeneous strategies were created by either replicating a single predator's learned policy across all three predators or by using a single fixed strategy across all predators. While performance for the learning predators was better in heterogeneous than homogeneous teams, performance for the analytical predators showed no advantage for heterogeneous teams.

[3] Here, we differentiate between strategy and policy by referring to policy as the instantaneous state-action mapping (i.e., given a state, an agent will take an action), whereas, strategy is the more general term that includes a temporal component which implies an inferred goal associated with the state-action mappings. Therefore, our conclusions are with respect to strategies.

4.1 Emergent Collaboration

Levels of performance vary among unique homogeneous sets of agents, as seen in Figure 3. We interpret this result to indicate that different learning agents conclusively learned different policies, which we also directly tie to learned strategies. Alternatively, it could be that the agents have learned the same general strategy with differential levels of effective execution. Either way, when combined in a heterogeneous team, these different strategies produce a collaborative effort that appears to be emergent, producing significantly better performance over a single replicated policy. Perhaps using replicated and unique strategies (e.g., combining 2 'Agent 0' policies with 1 'Agent 1' policy) can produce better performance than three different policies in the demonstrated paradigm. However, we are yet to investigate this. However, as was shown with the well-defined fixed strategy policies (i.e., Chasers and Interceptors), an arbitrary mixing of agent policies does not automatically result in better performance.

Ergodicity was demonstrated in prior work showing that learning agents appear to settle on the same pattern of movement regardless of episode length once a sufficient number of time steps have elapsed, and this pattern appears to be representative of the agent's policy (i.e., the state-action mapping that was learned through training). It would be of interest to see how the behavior patterns of trained agents change in various combinations of homogeneous and heterogeneous learned strategies (or policies), and further, how the insertion of fixed strategy agents impacts the behavior of trained agents, whether working collaboratively (as a fellow predator agent) or adversarially (as a prey agent).

4.2 Mixed Strategy Agents

While the heterogeneity did not improve the performance of analytical teams to anywhere near the performance of learning teams in the results presented above, heterogeneity within an agent through mixing well-defined strategies (e.g., Chaser and Interceptor) within a single agent's policy may produce a more sophisticated hybrid or mixed strategy agent against adversarial partners. This may be an agent that performs situational switching between well-defined strategies. It would be of interest to see if a group of 3 predator agents that learn to mix fixed strategies (i.e., adopt a policy that simply switches between analytically-defined behaviors) against a trained prey could perform better than the basic analytical agents described in this work.

Seeing as the trained prey was able to perform so well against the Chaser and Interceptor analytical agents, it may be that the trained predators at some point utilized similar strategies or some combination of them during learning. It could also be that the fixed strategies still exist in the learned policies but not in the simple form that we have presented. Finding some way to classify the behavior in the learned policies of the trained agents may allow us to extrapolate whether the agents' overall behavior is comprised of a mixing of well-defined strategies that are utilized when the appropriate conditions are met.

5 Conclusion

The results presented in this paper demonstrate that, while heterogeneity among teammates can improve team performance, heterogeneity by itself is not sufficient. Furthermore, even successful heterogeneous teams may not be a good solution. For example, the hard-coded diversity demonstrated by our fixed-strategy cases did not provide improved team performance over the trained agents. Extensions of this work may lead to successful, adaptive human teaming (treating the human as a heterogeneous strategy partner) that will need to be built upon trust and shared purpose. An approach to instantiate this resiliency in computational agents is to engender trust and shared purpose among their human collaborators by diversifying training (e.g., making learning agents robust to human complacent behavior [13]). Together, these lines of research push towards the design of optimally collaborative computational teammates that are not necessarily individually optimal agents, but work seamlessly with new partners whether computational or human.

References

1. Asher, D., Barton, S., Zaroukian, E., Waytowich, N.: Effect of cooperative team size on coordination in adaptive multi-agent systems. In: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, vol. 11006, p. 110060Z. International Society for Optics and Photonics (2019)
2. Asher, D.E., Zaroukian, E., Perelman, B., Perret, J., Fernandez, R., Hoffman, B., Rodriguez, S.S.: Multi-agent collaboration with ergodic spatial distributions. In: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II, vol. 11413, p. 114131N. International Society for Optics and Photonics (2020)
3. Barton, S.L., Waytowich, N.R., Zaroukian, E., Asher, D.E.: Measuring collaborative emergent behavior in multi-agent reinforcement learning. In: International Conference on Human Systems Engineering and Design: Future Trends and Applications, pp. 422–427. Springer (2018)
4. Barton, S.L., Zaroukian, E., Asher, D.E., Waytowich, N.R.: Evaluating the coordination of agents in multi-agent reinforcement learning. In: International Conference on Intelligent Human Systems Integration, pp. 765–770. Springer (2019)
5. Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W.: Openai gym. arXiv preprint arXiv:1606.01540 (2016)
6. Case, N.: How to become a centaur. Journal of Design and Science (2018)
7. Hicks, H.G., Gullett, C.R., Phillips, S.M., Slaughter, W.S.: Organizations: Theory and behavior. McGraw-Hill Companies (1975)
8. Humann, J., Jin, Y., Madni, A.M.: Scalability in self-organizing systems: An experimental case study on foraging systems. In: Disciplinary Convergence in Systems Engineering Research, pp. 543–557. Springer (2018)
9. Isaacs, R.: Differential Games. John Wiley and Sons, Inc. (1965)
10. Lowe, R., Wu, Y.I., Tamar, A., Harb, J., Abbeel, O.P., Mordatch, I.: Multi-agent actor-critic for mixed cooperative-competitive environments. In: Advances in neural information processing systems, pp. 6379–6390 (2017)

11. Oster, G.F., Wilson, E.O.: Caste and ecology in the social insects. Princeton University Press (1978)
12. Ramana, M.V., Kothari, M.: Pursuit-evasion games of high speed evader. Journal of intelligent & robotic systems **85**(2), 293–306 (2017)
13. Rodriguez, S.S., Chen, J., Deep, H., Lee, J., Asher, D.E., Zaroukian, E.: Measuring complacency in humans interacting with autonomous agents in a multi-agent system. In: T. Pham, L. Solomon, K. Rainey (eds.) Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II, vol. 11413, pp. 145 – 158. International Society for Optics and Photonics, SPIE (2020). DOI 10.1117/12.2559474. URL https://doi.org/10.1117/12.2559474
14. Shishika, D., Paulos, J., Dorothy, M.R., Hsieh, M.A., Kumar, V.: Team composition for perimeter defense with patrollers and defenders. In: 2019 IEEE 58th Conference on Decision and Control (CDC), pp. 7325–7332. IEEE (2019)
15. Song, Y., Kim, J.H., Shell, D.A.: Self-organized clustering of square objects by multiple robots. In: International Conference on Swarm Intelligence, pp. 308–315. Springer (2012)
16. Stone, P., Veloso, M.: Multiagent systems: A survey from a machine learning perspective. Autonomous Robots **8**(3), 345–383 (2000)
17. Tampuu, A., Matiisen, T., Kodelja, D., Kuzovkin, I., Korjus, K., Aru, J., Aru, J., Vicente, R.: Multiagent cooperation and competition with deep reinforcement learning. PloS one **12**(4) (2017)