Sharing is Not Needed: Modeling Animal Coordinated Hunting with Reinforcement Learning

Minglu Zhao[1] Ning Tang[1] Annya L. Dahmani[2] Yixin Zhu[1] Federico Rossano[3] Tao Gao[1,4]

***Abstract*— Coordinated hunting is widely observed in animals, and sharing rewards is often considered a major incentive for this success. While most results on this topic are correlational, we reveal the causal roles of sharing rewards through computational modeling with a state-of-the-art Multi-agent Reinforcement Learning (MARL) algorithm. Using a coordinated hunting task with a group of predators hunting one prey, we show that sharing rewards is *neither necessary nor sufficient* for modeling animal coordinated hunting. Hunting coordination modeled through sharing rewards 1) suffers from the free-rider problem, 2) plateaus at a small group size, and 3) is not a Nash equilibrium. Moreover, individually rewarded predators outperform predators that share rewards, especially when the hunting is difficult, the group size is large, and the action cost is high. We conclude that animal coordinated hunting can be successfully modeled through reinforcement learning only when the agents are selfish, and not when the rewards are shared. [1].Our results further offer computational support to the explanation of chimpanzee behavior that agents with only selfish interests can form coordinated hunting, and sharing rewards might simply be a byproduct of hunting, instead of an intelligent design to facilitate coordination [1].**

I. Introduction

Coordinated hunting has been broadly observed in the animal kingdom for many different species such as wolves [2], hyenas [3], dolphins [4], ravens [5], and hawks [6], whereas the majority of in-depth discussion on coordination mechanisms focuses on chimpanzee behavior. Chimpanzees hunt for meat in all known populations, with the red colobus monkeys being the primary prey where both species exist [7]. Anthropological studies based on field observations suggest that chimpanzees exhibit sophisticated human-like cooperation, such as playing complementary roles during hunting, which includes drivers, blockers, chasers, and ambushers [8]. Consequently, understanding the motivation of such coordinated behavior provides insights on the evolutionary history of human cooperation.

Sharing rewards has been considered as a major incentive for animals' success in coordinated hunting, especially for chimpanzees. It seems to encourage participation in group hunting, which leads to higher hunting success [9]. Moreover, evidence has shown that sharing rewards further contributes to chimpanzee bonding through reciprocity [10], [11], reducing begging harassment [12], [13], exchanging meat for sex [11], and securing dominance [11].

[1] Department of Statistics, University of California, Los Angeles, CA 90025, USA. Emails: minglu.zhao@ucla.edu, ningtangcog@gmail.com, yixin.zhu@ucla.edu, tao.gao@stat.ucla.edu,

[2]Department of Psychology, UCLA. adahmani@g.ucla.edu

[3]Department of Cognitive Science, University of California, San Diego, CA 92093, USA. frossano@ucsd.edu

[4]Department of Communication, UCLA.

[1]See videos and supplementary material on project website https://sites.google.com/view/icra2021ws-sharingrewards.

However, since existing animal studies are mostly observational, they can only indicate a correlation, while the causal effects of sharing rewards on coordination remain unclear due to the lack of causal evidence from formal experimental manipulations. In fact, it has been argued that sharing rewards is neither necessary nor sufficient for coordinating animal group behavior. It **could be unnecessary**, since chimpanzee coordinated hunting may not be based on any type of sharedness at all, but is mainly driven by selfish interests [1]. In such a case, during hunting, each individual chimpanzee simply takes actions to maximize its self-interest based on other agents' locations. Sharing rewards **could be insufficient** to support coordinated behavior. One critical challenge for coordination is the free-rider problem [14]: Rational individuals have little incentive to contribute to the production of a common good, given the costs they would incur, since they will benefit from the shared good whether or not they contribute. However, the free-rider problem has not been highlighted in existing observational studies of animal hunting. Theories of human cooperation suggest that other cognitive infrastructures are necessary to solve this issue, including cheater detection and punishment [15], commitment [16], fairness [17], and accountability [18]. Evoking these complex normative and moral concepts requires a stronger definition of cooperation beyond sharing rewards [19].

With the aforementioned conflicting evidence and viewpoints from observational studies, here we study the causal effects of sharing rewards on the performance of coordinated hunting from a modeling perspective. There has been a long history of studies in the ultimate mechanisms of cooperation through an evolutionary perspective, which focuses on consequences at the population level [20], [21], [22]. Here, we focus on the proximate mechanisms [23], [24] of individual agents' decision-making. We use a cognitively realistic artificial intelligence model, Reinforcement Learning (RL), in which an agent aims to maximize its accumulated, long-term rewards by learning how to act from trial and error [25]. RL is a prominent model for animal learning with deep roots in psychology and neuroscience [26]. RL is also a state-of-the-art artificial intelligence model, which, combined with deep neural networks [27], is able to generate complex intelligent behaviors, reaching human-expert level performance in games like Atari [28] and Go [29], [30].

More critically, Multi-agent Reinforcement Learning (MARL), as an extension of RL, has been successfully applied to challenging group coordination scenarios, such as autonomous-driving coordination [31], teaming in Dota 2 [32] and StarCraft [33]. MARL offers a generic solution to different applications simply through adjusting the relationship between agents' reward functions. For competition,

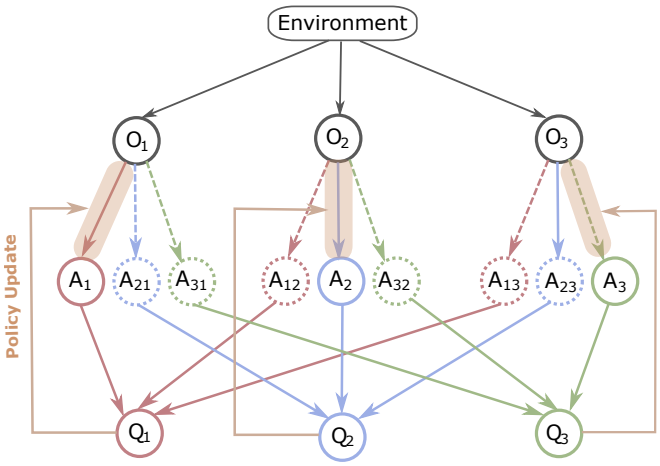
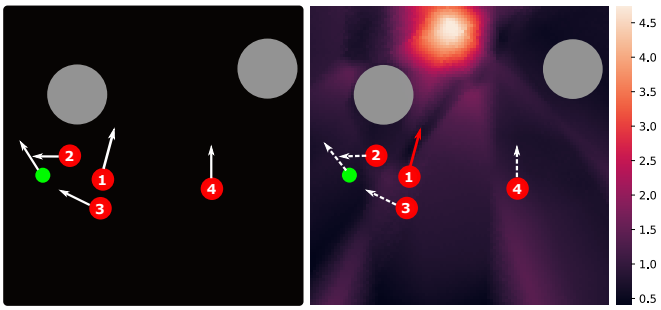


Fig. 1: **Illustration of multi-agent deep deterministic policy gradient (MADDPG) algorithm with three agents.** The MADDPG algorithm incorporates *centralized* evaluation and *decentralized* acting: Each agent i only has access to its own observations O_i to choose action A_i (policy indicated in the shaded area). When evaluating its action, each agent incorporates all other agents' observations to predict their actions (represented by dashed circles: A_{ij} denotes agent i's prediction of agent j's action-to-take) and then use the observations and action predictions to form a centralized evaluation Q_i for its own action. Agents would then update their policies to output actions that generate improved values.



(a) (b)

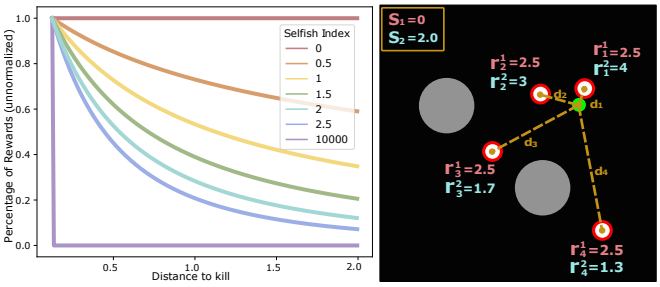
Fig. 2: **An illustration of the coordinated hunting task with a successful coordination policy.** Red circles represent the predators, and the green circle represents the prey. Two grey circles represent the obstacles that agents cannot move through, which are also randomly located and remain stationary. a) One frame of agents' motions during coordination. Predator 1 in this case chooses to move upwards, potentially to block the prey's movement (as a "blocker"), instead of directly towards the prey (as a "chaser"), a choice that indicates a sophisticated coordination strategy incorporated by the agents. b) Value landscape of predator 1's potential positions. During the evaluation phase of predator 1, it makes predictions on others' actions-to-take (indicated by dashed arrows). It then uses the predictions, together with other agents' observations, to evaluate the value of its own action. Here, we plot the value of different positions of predator 1 generated from its model. The predator's policy encourages itself to move towards the state that induces a greater value.

the reward functions are zero-sum. Critically, MARL defines cooperation as agents aligning their rewards through the same reward function [34], effectively splitting the group reward among all coordinating agents. With the critical position taken by reward-sharing in MARL, it is theoretically important to reveal the **causal** role of reward distribution in generating coordinated hunting with this model.

One particular algorithm in MARL, multi-agent deep deterministic policy gradient (MADDPG) [35], has been successfully applied to a multi-agent hunting game through agents sharing rewards. It shows that a group of predators can learn from scratch to coordinate the hunting of an intelligent prey. The algorithm is decentralized at the top level, with each agent learning its own model, instead of having a unified policy copied for all predators (Fig. 1). The training is cognitively intelligent in two ways. First, it involves cognitive constraints: when taking actions, each agent can only refer to its own observation, without accessing observations from others. Second, each agent treats others as actual agents, instead of random objects in the environment, and predicts what they will do next, a process that can be interpreted as a primitive version of Theory of Mind [36]. Moreover, an agent's evaluation of an action is based on all agents' states and predicted actions, echoing Tomasello's theory of coordinated hunting [1]. Agents will then update their policies to output actions that would improve this evaluation. The planning remains individualized because agents only care about their own actions while evaluating the situation from a holistic perspective.

However, as the focus of the MADDPG study is not in explaining realistic animal behavior, there are critical artificial components that make the conclusion not generalizable to animal hunting. First, the framework takes sharing rewards as an assumption and does not provide a comparison with cases using individual rewards, which fails to provide causal evidence for the effect of reward distribution. Second, the predators and prey have no action cost in the environment; thus, the free-rider problem is avoided altogether, since the only motivation for free-riding is to avoid individual costs in cooperation. Third, to achieve better results in training, the environment rewards predator agents for "bites," instead of "kills," to create frequent reward signals, which helps with the model training. However, such a setting is unrealistic and even opposite in real-world hunting, where only kills matter and can provide substantial material rewards. Bites without kills may in fact provide a negative reward to the predator, as it introduces chances of injuries and costs efforts. As such, it remains nebulous whether the computational model can indeed handle scenarios with only kill signals.

To test the causal effects of sharing rewards in modeling animal coordinated hunting, we adopt a coordinated hunting game setting [35] by populating the environment with multiple predators and one prey (Fig. 2). Predators are rewarded *only* after killing the prey, which happens in 20 percent of biting instances; a bite occurs when a predator collides with the prey. After killing the prey, a constant reward will be allocated to predators based on the reward mechanism in the specific experimental condition. Predators have individual action costs proportional to the force they exert. Preys are trained with the same algorithm with a negative reward at each bite or kill.



(a) (b)

Fig. 3: **Illustrations of the reward distribution.** a) Reward distribution as an exponential function of distance-to-kill. Percentage of rewards obtained (unnormalized in the figure) decreases as an agent's distance-to-kill increases. We chose seven values of selfish indices to systematically test the effects of reward distributions. b) An example scenario showing rewards to be obtained by four agents under two example selfish indices, $s_1 = 0, s_2 = 2$ (red and turquoise color, respectively). Agents split the reward evenly when $s = 0$, and are sensitive to the distance-to-kill at $s = 2$. Agent i's reward obtained is represented by r_i^s under selfish index s, which is calculated through the agent's distance-to-kill, d_i.

II. Experimental Design

We systematically test MADDPG's performance in coordinated hunting with experimental manipulations inspired by anthropological and animal studies.

Reward distribution among predators: Anthropological observations indicate that proximity to prey at the moment it was killed is an essential factor when chimpanzees decide how to split the spoil [13]. Here we control the reward distribution among predators as a function of the distance-to-kill. Sensitivity to the distance-to-kill illustrates a selfish index. With a high selfish index, the reward distribution concentrates on the predators close to the kill. When the predators are purely selfish (*i.e.*, with an infinite selfish index), after one predator kills the prey, it will only reward itself since it is the one closest to the prey. With a low selfish index, rewards will be broadly dispersed. When the predators are purely unselfish (*i.e.*, with a zero selfish index), rewards will be evenly distributed, regardless of agents' distance-to-kill. All predators follow the same mechanism in one condition. Formally, we define the reward distribution as an exponential function of the distance-to-kill, such that

$$R_i \propto (d_i + 1 - k)^{-s}, \quad (1)$$

where an agent i with d_i distance-to-kill receives R_i proportion of reward, with selfish index s. The constant k denotes the minimum distance between two agents; see Fig. 3.

Action cost for testing the free-rider problem: Agents are motivated to free-ride in coordinated hunting only to avoid the individual costs [14]. As the action cost increases, agents would prefer to stay static to reduce individual action costs while at the same time obtain allocated rewards. Accordingly, to test the severity of free-rider problems as a function of the reward distribution, we define the action costs to be proportional to the force exerted by agents, with action cost for agent i, $C_i = a * F_i$, where F_i is the force exerted by agent i, and a denotes the action cost ratio in the specific condition. The action costs are applied to individual agents no matter which reward mechanism they take.

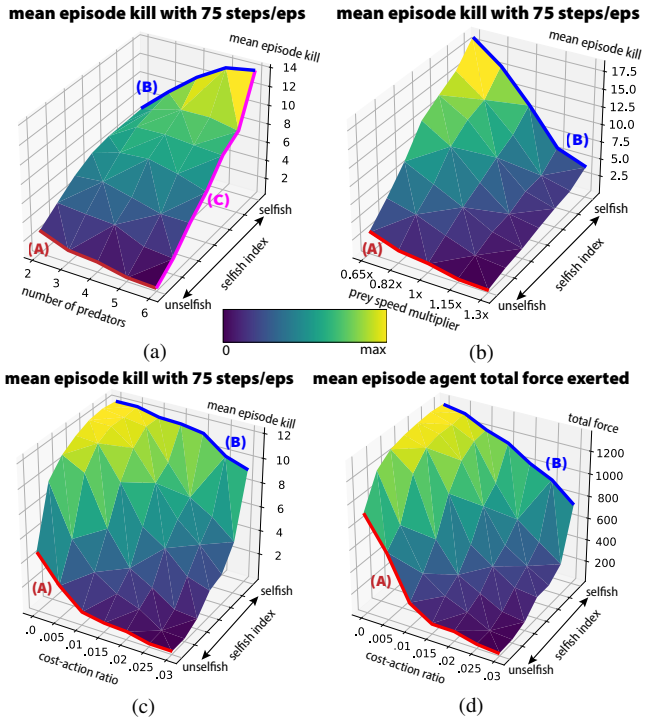


Fig. 4: **Modeling results visualized as landscapes spanned by different variables.** Brighter color (yellow) denotes larger values, and darker color (blue) smaller values. a) The performance of selfish agents increases linearly with the group size (linear regression, p $<$ .005, line (B)). Unselfish agents' performance remains the same or even drops when having a larger group (line (A), p = 0.013). Taking the group size of 6 as an example (line (c)), without loss of generality, the more selfish the predators, the better their performance (p $<$ .001). b) The predators' performance decreases under all reward mechanisms, as the speed of prey increases. c) The performance of all agents decreases as the action cost increases (p $<$ .001). d) More selfish agents have their action force less sensitive to action costs (unselfish agents indicated by line (A), and selfish agents indicated by line (B)).

Group size: Evidence in animal studies has shown that hunting party size is positively correlated with hunting success for many different species [37], [9], [2], [6], [38]. Here we test how the reward distribution interacts with group size by having different numbers of predators in the group.

Hunting risk: Hunting risks have been a significant factor influencing the hunting behavior of many species. Wolves display a higher level of participation in riskier hunting [2]. Chimpanzee hunting has a low success rate and thus renders hunting an unnecessary activity for some groups, and some choose to hunt only in a situation of full nutritional abundance [7]. To evaluate the interaction between reward distribution and hunting risks, we manipulate hunting risks by the speed of prey as compared to the predator.

III. Results

Our results indicate that there are significant main effects for all four variables, and the selfish index is significantly

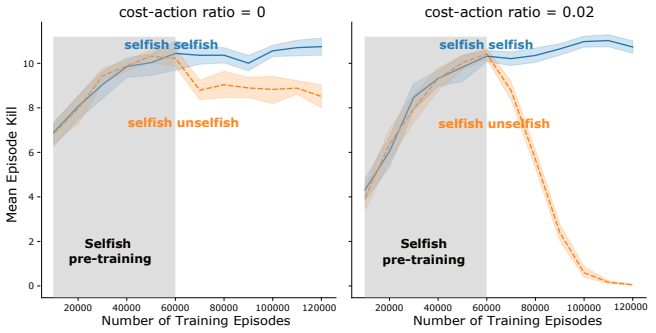


Fig. 5: **Model performance in equilibrium testing.** Selfish agents are further trained by sharing or individual reward mechanisms. Performance of pretrained agents significantly decreased after training with shared rewards ($p < 0.001$, df = 18 for both tests).

intertwined with all three other variables (second-order multiple regression model, $p < .001$ for all terms). In our further analysis, we investigate two-way interaction terms by aggregating the other two conditions; see Fig. 4. Our results indicate that performance (*i.e.*, the number of kills in one episode) of selfish agents increases linearly with the group size, while unselfish agents' performance remains the same or even drops when having a larger group. The more selfish the predators, the better their performance. The change is monotonic, with the most selfish predators achieving the best performance; see Fig. 4a. The performance of all agents decreases as the action cost increases, and increasing action costs hurts unselfish agents more than selfish agents—unselfish agents fail to obtain any rewards as the action cost reaches .01 level, whereas selfish agents still maintain a high level of performance even under the largest action cost condition; see Fig. 4c. Furthermore, more selfish agents have their action force less sensitive to action costs. The most unselfish agents decide almost not to move at all when there is a small action cost; see Fig. 4d. *Such a result strongly indicates the presence of the free-rider problem under the reward-sharing mechanism.* Lastly, as the speed of prey increases, the predators' performance decreases under all reward mechanisms, with the most selfish agents performing the best in all conditions; see Fig. 4b.

Equilibrium testing: Having shown that sharing rewards cannot generate robust coordination through learning, our further experiments focus on whether, given a successful coordinated policy, coordination can be maintained under sharing rewards, which is the same as testing whether successful coordination is a Nash Equilibrium through sharing rewards. We adopt models of already-coordinated selfish agents and train them for another around with both the individual-reward strategy and the reward-sharing strategy. Our results indicate that performance of agents pretrained with individual rewards for 60,000 episodes significantly decreased after training with shared rewards for another 60,000 episodes; see Fig. 5. We conclude that successful coordination through sharing rewards is not an equilibrium, since all agents' policies deviate from it in further training.

Discussion

Our results suggest that sharing rewards is neither necessary nor sufficient for modeling animal coordinated hunting behavior with reinforcement learning. It is unnecessary since models without any sharedness (selfish agents) achieve good training results in the environment and even outperform agents that share rewards (unselfish agents). It is insufficient for three reasons. First, our results indicate a free-rider problem for unselfish agents. Specifically, when agents share rewards and have their movements subject to individual action costs, they become reluctant to move, which negatively affects the group's performance. Second, unselfish agents' hunting performance does not improve when the group size increases, which contrasts with the observational evidence that hunting success should be positively correlated with the group size. Third, the reward-sharing mechanism cannot maintain a coordinated performance, with agents' actions deviated from a well-trained policy, possibly due to the free-rider problem. Hence, our results support Tomasello's theory of chimpanzee behavior that agents with selfish interests are capable of forming successful coordinated hunting [1]. Furthermore, chimpanzees would participate in group hunting due to selfish motivations, not to expectations about the sharing of rewards. Sharing rewards, in this way, might simply be a byproduct of chimpanzee hunting, instead of an intelligent design or the cause that improves the coordination performance, since the hunting performance deteriorates as the reward distribution gets more distributed.

MARL has been taken as a competitive model of cooperation through agents sharing rewards. However, our results indicate that this mechanism is not a required precondition for generating coordinated behavior and might even produce worse performance than training without such assumption. Suffering from the free-rider problem, coordination generated by multiple agents sharing rewards is indeed a special case when the action cost is zero. Moreover, while various applications of RL algorithms are realized through combining the generic algorithm with adjusting reward functions, we believe that this should not be the whole story when modeling human-like cooperation. Evidence and theories in comparative psychology suggest that cooperation is qualitatively different from animal coordination [39]. In this case, although selfish agents in our modeling achieve better performance than unselfish agents, the selfish strategy alone is far from sufficient to accomplish most cooperation tasks human beings face. To better model human cooperation, certain types of sharedness beyond purely sharing rewards are indispensable, and a more sophisticated mechanism involving the idea of the shared agency may be required [40], [39].

Appendix

a) MADDPG: Our implementation of the MADDPG algorithm mostly follows the original implementation [35]. Each group of models (including the predators and prey) is trained for 60,000 episodes, with 75 time steps in each episode. We use Adam optimizer [41] with a learning rate of .01, the soft update rate τ of .01, discount factor γ of .95. Policies are parameterized by a two-layer ReLU MLP with 128 units per layer for both predators and the prey. Memory buffer size is 10^6 with a mini-batch size of 1024.

b) Evaluation: During evaluation, we use the same environment and record the number of kills by each group within an episode to represent the group's performance.

To account for the possible inconsistency in the preys' performance and avoid models' overfitting issues, for each sampled trajectory, we have the predators in the specific condition chase a prey that is randomly selected from all trained prey models.

REFERENCES

[1] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll, "Understanding and sharing intentions: The origins of cultural cognition," *Behavioral and Brain Sciences*, vol. 28, no. 5, pp. 675–691, 2005.

[2] D. R. MacNulty, A. Tallian, D. R. Stahler, and D. W. Smith, "Influence of group size on the success of wolves hunting bison," *PloS One*, vol. 9, no. 11, p. e112884, 2014.

[3] K. E. Holekamp, L. Smale, R. Berg, and S. M. Cooper, "Hunting rates and hunting success in the spotted hyena (crocuta crocuta)," *Journal of Zoology*, vol. 242, no. 1, pp. 1–15, 1997.

[4] S. K. Gazda, R. C. Connor, R. K. Edgar, and F. Cox, "A division of labour with role specialization in group–hunting bottlenose dolphins (tursiops truncatus) off cedar key, florida," *Proceedings of the Royal Society B: Biological Sciences*, vol. 272, no. 1559, pp. 135–140, 2005.

[5] R. Yosef and N. Yosef, "Cooperative hunting in brown-necked raven (corvus rufficollis) on egyptian mastigure (uromastyx aegyptius)," *Journal of ethology*, vol. 28, no. 2, pp. 385–388, 2010.

[6] J. C. Bednarz, "Cooperative hunting harris' hawks (parabuteo unicinctus)," *Science*, vol. 239, no. 4847, pp. 1525–1527, 1988.

[7] N. E. Newton-Fisher, *Chimpanzee hunting behavior*. Springer-Verlag, 2007.

[8] C. Boesch, "Joint cooperative hunting among wild chimpanzees: Taking natural observations seriously," *Behavioral and Brain Sciences*, vol. 28, no. 5, pp. 692–692, 2005.

[9] L. Samuni, A. Preis, A. Mielke, T. Deschner, R. M. Wittig, and C. Crockford, "Social bonds facilitate cooperative resource sharing in wild chimpanzees," *Proceedings of the Royal Society B*, vol. 285, no. 1888, p. 20181643, 2018.

[10] J. C. Mitani and D. P. Watts, "Why do chimpanzees hunt and share meat?" *Animal Behaviour*, vol. 61, no. 5, pp. 915–924, 2001.

[11] T. Nishida, T. Hasegawa, H. Hayaki, Y. Takahata, and S. Uehara, "Meat-sharing as a coalition strategy by an alpha male chimpanzee," *Topics in primatology*, vol. 1, pp. 159–174, 1992.

[12] I. C. Gilby, "Meat sharing among the gombe chimpanzees: harassment and reciprocal exchange," *Animal Behaviour*, vol. 71, no. 4, pp. 953–963, 2006.

[13] M. John, S. Duguid, M. Tomasello, and A. P. Melis, "How chimpanzees (pan troglodytes) share the spoils with collaborators and bystanders," *PloS One*, vol. 14, no. 9, p. e0222795, 2019.

[14] M. Olson, *Logic of collective action: Public goods and the theory of groups (Harvard economic studies. v. 124)*. Harvard University Press, 1965.

[15] T. H. Clutton-Brock and G. A. Parker, "Punishment in animal societies," *Nature*, vol. 373, no. 6511, pp. 209–216, 1995.

[16] M. Gilbert, "Obligation and joint commitment," *Utilitas*, vol. 11, no. 2, pp. 143–163, 1999.

[17] R. J. Arneson, "The principle of fairness and free-rider problems," *Ethics*, vol. 92, no. 4, pp. 616–633, 1982.

[18] D. De Cremer and M. Barker, "Accountability and cooperation in social dilemmas: The influence of others' reputational concerns," *Current Psychology*, vol. 22, no. 2, pp. 155–163, 2003.

[19] M. Tomasello, *Why we cooperate*. MIT press, 2009.

[20] R. Boyd, H. Gintis, S. Bowles, and P. J. Richerson, "The evolution of altruistic punishment," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 100, no. 6, pp. 3531–3535, 2003.

[21] M. A. Nowak, "Five rules for the evolution of cooperation," *Science*, vol. 314, no. 5805, pp. 1560–1563, 2006.

[22] J. Gross and C. K. De Dreu, "The rise and fall of cooperation through reputation and group polarization," *Nature Communications*, vol. 10, no. 1, pp. 1–10, 2019.

[23] E. Mayr, "Cause and effect in biology," *Science*, vol. 134, no. 3489, pp. 1501–1506, 1961.

[24] N. Tinbergen, "On aims and methods of ethology," *Zeitschrift für tierpsychologie*, vol. 20, no. 4, pp. 410–433, 1963.

[25] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[26] E. O. Neftci and B. B. Averbeck, "Reinforcement learning in artificial and biological systems," *Nature Machine Intelligence*, vol. 1, no. 3, pp. 133–143, 2019.

[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. [Online]. Available: https://doi.org/10.1038/nature14539

[28] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[29] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[30] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[31] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "Safe, multi-agent, reinforcement learning for autonomous driving," *arXiv preprint arXiv:1610.03295*, 2016.

[32] C. Berner, G. Brockman, B. Chan, V. Cheung, P. Debiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, *et al.*, "Dota 2 with large scale deep reinforcement learning," *arXiv preprint arXiv:1912.06680*, 2019.

[33] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, "Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4295–4304.

[34] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 2, pp. 156–172, 2008.

[35] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[36] H. M. Wellman, *The child's theory of mind*. The MIT Press, 1992.

[37] J. C. Mitani and D. P. Watts, "Demographic influences on the hunting behavior of chimpanzees," *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, vol. 109, no. 4, pp. 439–454, 1999.

[38] S. Creel and N. M. Creel, "Communal hunting and pack size in african wild dogs, lycaon pictus," *Animal Behaviour*, vol. 50, no. 5, pp. 1325–1339, 1995.

[39] M. Tomasello, *Origins of human communication*. MIT press, 2010.

[40] M. Kleiman-Weiner, M. K. Ho, J. L. Austerweil, M. L. Littman, and J. B. Tenenbaum, "Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2016.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.