Multi-Agent Curricula and Emergent Implicit Signaling

Niko A. Grupen
Cornell University
Ithaca, New York, USA

Daniel D. Lee
Cornell Tech
New York, New York, USA

Bart Selman
Cornell University
Ithaca, New York, USA

ABSTRACT

Emergent communication has made strides towards learning communication from scratch, but has focused primarily on protocols that resemble human language. In nature, multi-agent cooperation gives rise to a wide range of communication that varies in structure and complexity. In this work, we recognize the full spectrum of communication that exists in nature and propose studying lower-level communication. Specifically, we study emergent implicit signaling in the context of decentralized multi-agent learning in difficult, sparse reward environments. However, learning to coordinate in such environments is challenging. We propose a curriculum-driven strategy that combines: (i) velocity-based environment shaping, tailored to the skill level of the multi-agent team; and (ii) a behavioral curriculum that helps agents learn successful single-agent behaviors as a precursor to learning multi-agent behaviors. Pursuit-evasion experiments show that our approach learns effective coordination, significantly outperforming sophisticated analytical and learned policies. Our method completes the pursuit-evasion task even when pursuers move at half of the evader's speed, whereas the highest-performing baseline fails at 80% of the evader's speed. Moreover, we examine the use of implicit signals in coordination through position-based social influence. We show that pursuers trained with our strategy exchange more than twice as much information (in bits) than baseline methods, indicating that our method has learned, and relies heavily on, the exchange of implicit signals.
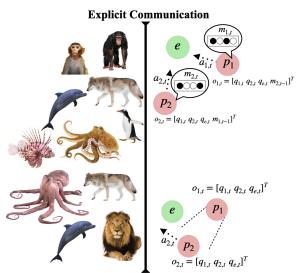
KEYWORDS

Multi-Agent Systems; Reinforcement Learning; Communication

1 INTRODUCTION

Communication is a critical scaffolding for coordination. It enables humans and animals alike to coordinate on complex tasks, synchronize plans, allocate team resources, and share missing state. Understanding the process through which communication emerges has long been a goal of philosophy, linguistics, cognitive science, and AI. Recently, advances in multi-agent reinforcement learning (MARL) have propelled computational studies of emergent communication that examine the representations and social conditions necessary for communication to emerge in situated multi-agent populations [39]. Existing approaches, targeting language-like communication, have shown that it is possible for agents to learn protocols that

Figure 1: Left: Animal communication spans a complete spectrum from explicit to implicit communication. Right: A pursuit-evasion game with two pursuers p_1 and p_2 and an evader e (with positions $q_{1,t}$, $q_{2,t}$, and $q_{e,t}$ at time t). Emergent communication only considers language-like communication (top), where agents share information over specialized sender-receiver architectures—e.g. p_1 receives the extended observation $o_{1,t} = [q_{1,t}\ q_{2,t}\ q_{e,t}\ m_{2,t-1}]^T$ and selects actions using the extra information $m_{2,t-1}$ that p_2 chose to share at the previous time-step (and vice versa for p_2). With implicit signals (bottom), p_1 selects a movement action $a_{1,t}$ based only on the observation $o_{1,t} = [q_{1,t}\ q_{2,t}\ q_{e,t}]^T$. The only information p_1 gets from p_2 is generated by p_2's physical position (and vice versa for p_2). We posit that implicit signals are an important step towards learning richer communication.

exhibit language-like properties such as compositionality [10, 58] and Zipf's Law [11] when given additional learning biases [23].

Though emergent communication is fundamentally an *ab initio* approach compared to top-down language learning [66], most recent methods at the intersection of emergent communication and multi-agent reinforcement learning have targeted protocols with complex structure and representational capacity, like that of human language [39, 43]. Such techniques equip agents with specialized sender-receiver architectures and allow them to exchange bits of information (or continuous signals) through these channels. One notable exception is the work of Jaques et al. [36], which considers how influential agents can pass information to other agents through their actions in the environment. We build on these ideas, recognizing the full spectrum of communication that exists in nature and

proposing the study of lower-level forms of communication that do not require specialized architectures as a first step towards more sophisticated communication (see Figure 1).

Multi-agent cooperation in nature yields a wide range of communication protocols that vary in structure and complexity. In animal communication [6], for example, reef-dwelling fish use body shakes [8, 65] and octopuses punch collaborators [60]—forms of non-verbal communication—whereas chimps [5], macaques [45], and gentoo penguins [13] each maintain a diverse vocal repertoire. Importantly, animals also exchange information through non-explicit channels. In the Serengeti, for example, lions use the posture of fellow pride members to stalk prey covertly [61]. Moreover, wolves [53] and dolphins [56], both frequent vocal communicators, communicate implicitly during foraging—adjusting their group formation based on the position and orientation of other pack members [32]. Breazeal et al. [7] define such forms of communication, called *implicit communication*, as "conveying information that is inherent in behavior but which is not deliberately communicated". Studies have shown that implicit cues are used frequently by humans as well. Human teams rely on gaze, facial expressions, and non-symbolic movement during cooperative tasks [24] including collaborative design [70], crowdsourcing [72], and multiplayer games (e.g. Hanabi [2, 40]).

The role of implicit communication in teamwork has been studied extensively in the human-computer [40, 62] and human-robot interaction [7, 9, 12, 37] communities, and even in multi-agent systems [29, 50]; but is less well-understood in the context of emergent agent-agent interactions [19]. This is due in large part to the anthropomorphized nature of standard definitions of implicit communication. Note that the key difference between explicit communication (a deliberate information exchange) and implicit communication (a non-deliberate exchange) in the definition above is *intent* [7, 29]. Understanding implicit communication in the context of artificial agents requires (i) ascribing human-like mental states to artificial agents and (ii) giving agents the ability to reason about each other's mental states with respect to a common goal (*à la* Cohen and Levesque [16]); which is a highly non-trivial undertaking.

In this work, we take a first step in this direction by examining the extent to which implicit communication aides coordination in multi-agent populations. Following the work of De Greeff and Nolfi [19] in evolutionary settings, we study *implicit signals*: signals generated by the physical position of an agent as it interacts with its environment. Implicit signals are observed passively by an agent's teammates, and so require neither special sender-receiver architectures nor reasoning about an agent's mental states. For RL agents, we distinguish implicit signals from explicit communication through each agent's actions and observations. As a simple example, consider the environment in Figure 1 where, at time t, agents p_1, p_2, and e are described by their positions $q_{1,t}$, $q_{2,t}$, and $q_{e,t}$ respectively. With implicit signals, p_1 selects a movement action $a_{1,t}$ based on the observation $o_{1,t} = [q_{1,t} \; q_{2,t} \; q_{e,t}]^T$—i.e. the only information p_1 gets from p_2 is the information contained in p_2's physical position; and vice versa. With explicit communication, p_1 emits a communicative action $m_{1,t}$ in addition to $a_{1,t}$ (and same for p_2). In this case, p_1 receives the extended observation $o_{1,t} = [q_{1,t} \; q_{2,t} \; q_{e,t} \; m_{2,t-1}]^T$ and can select actions using the extra information $m_{2,t-1}$ that p_2 chose to share at the previous time-step.

Experimentally, we study coordination and the role of implicit communication with pursuit-evasion games [33]. Pursuit-evasion simulates an important coordination task, foraging, where communication is especially impactful. To accurately model the conditions under which coordination and communication emerges in the animal kingdom, we prioritize decentralized multi-agent learning. Despite the challenges posed by decentralized approaches [26], this feature is critical in the emergence of animal communications and so we focus on this important class of multi-agent learning frameworks to exploit inspiration from the natural world. Moreover, to encourage teamwork, we target pursuit-evasion games in which the pursuers are slower than the prey (coordination is not needed with the pursuers have a velocity advantage). Similar to Lowe et al. [44], we study pursuit-evasion as a MARL problem.

We show that, under these conditions, naive decentralized learning fails to solve the pursuit-evasion task. There are a number of challenges that lead to this outcome. First, decentralized learning is non-stationary [44]—with multiple agents learning simultaneously, the learning problem is less stable for each agent individually. Next, since we prioritize tasks that require teamwork (e.g. pursuer speed less than evader speed), learning from scratch is difficult, as agents are forced to learn how to interact with the environment individually while simultaneously learning to coordinate with their teammates. Finally, pursuit-evasion is a sparse reward task—pursuers receive reward only when capturing the evader. During the early stages of training, when each agent is following a randomly initialized policy, it is virtually impossible for the agents to coordinate effectively enough to experience positive reward.

To address these challenges, we introduce a curriculum-driven approach for solving difficult, sparse reward coordination tasks. Our strategy, combining ideas from the ecological reinforcement learning [15] and automatic curriculum learning [55] literature, is motivated by the following insights: (i) we can create a *curriculum over task difficulty by shaping the environment* to the current skill level of the cooperative agents; (ii) we can aide agents in learning multi-agent team behaviors by forcing them to *learn single-agent behaviors first*. More specifically, we first adopt an environment shaping curriculum using velocity bounds, which allows pursuers to gather initial experience at a velocity greater than that of the evader, and then tune their strategies over time as velocity decreases. Then, we introduce a curriculum over agent behavior that warm-starts cooperative multi-agent learning by first seeding each agent with experience of successful single-agent behaviors.

Our first result confirms the importance of curriculum-driven multi-agent learning. We compare the performance of our strategy to ablations over the multi-agent curricula individually and show that, using our strategy, decentralized agents learn to solve difficult coordination tasks, such as the pursuit-evasion game.

Next, we stress-test both the strength of coordination and capacity for implicit communication of policies learned with our method. We compare the performance of multi-agent coordination learned with our method against a set of analytical and learned strategies that represent ablations over both coordination and implicit communication—i.e. outperforming one of these methods is equivalent to surpassing that method's level of sophistication in coordination, capacity for implicit communication, or both.

Empirical results show that our method significantly outperforms highly sophisticated multi-agent strategies. Coordination strategies learned with our method successfully complete the pursuit-evasion task at a **speed ratio of 0.5** (i.e. the pursuers are moving at half the evader's speed), whereas each of the baseline methods fails to complete the pursuit-evasion task when pursuer speed is only slightly below evader speed (**speed ratio of 0.8**).

Further, we hone in on the nature of each strategy's coordination more directly by examining agent behavior during the most important subsets of the trajectory—those that immediately precede the accumulation of reward. In the context of pursuit-evasion, we examine the distribution of pursuer locations relative the evader at the time of capture. By comparing rotational symmetry and rotational invariance in the agents' capture distributions, we find that our strategy learns structured coordination while simultaneously allowing pursuers to make dynamic adjustments to their position relative the evader to successfully achieve capture.

Finally, we examine the role of implicit signals (as introduced by De Greeff and Nolfi [19]) more directly using measures of social influence. Influence measures such as Instantaneous Coordination [36] quantify the extent to which one agent's behavior influences its teammates. We repurpose this method to measure the exchange of implicit signals as *position-based social influence*. Results of this study show that pursuers trained with our strategy exchange **up to 0.375 bits of information per time-step** on average, compared to a maximum of **only 0.15 bits on average** from the baseline methods. This indicates that our method learns, and relies heavily on, the exchange of implicit signals.

Preview of contributions Altogether, our work is summarized by the following contributions:

- We highlight the importance of studying implicit signals as a form of emergent communication.
- We introduce a curriculum-driven learning strategy that enables cooperative agents to solve difficult coordination tasks with sparse reward.
- We show that, using our strategy, pursuers learn to coordinate to capture a superior evader, significantly outperforming sophisticated analytical and learned methods. Coordination strategies learned with our method complete the pursuit-evasion task at a speed ratio of 0.5 (i.e. half the evader's speed), whereas each of the competing methods fails to complete the pursuit-evasion task at any speed ratio below 0.8.
- We examine the use of implicit signals in emergent multi-agent coordination through position-based social influence. We find that pursuers trained with our strategy exchange up to 0.375 bits of information per time-step on average, compared to a maximum of only 0.15 bits on average from competing methods, indicating that our method has learned, and relies heavily on, the exchange of implicit signals.

2 RELATED WORK

2.1 Emergent Communication

Emergent communication examines the process by which cooperative agents learn communication protocols as a means to completing difficult multi-agent tasks [39]. Recent work has shown that

MARL agents converge upon useful communication in referential games [31] and can even develop language-like properties such as compositionality [10, 58] and Zipf's Law [11] when exposed to additional structural learning biases. More recently, this class of techniques has expanded to include complex situated environments [18], high-dimensional observations [17], and the negotiation of belief states [28]. Further work has shown that influential communication can be incentivized through additional learning objectives (i.e. inductive biases) [23] and reward shaping [36].

Implicit interactions are less well-studied in the emergent communication literature, despite their importance to teamwork, as outlined in the human-computer [40, 62] and human-robot interaction [7, 9, 12, 37] communities. Implicit communication has been studied for cooperative multi-agent tasks [29, 30, 50], though not in the context of emergent behavior. Though some studies have shown that agents can learn to communicate non-verbally through actions and gestures [47], such forms of action-space communication [1] are examples of non-verbal explicit communication, whereas our goal is to study completely implicit signals. Most similar to our work is that of De Greeff and Nolfi [19], which examines both implicit and explicit communication in evolutionary settings. Our work can be interpreted as bridging these ideas with the MARL literature.

2.2 Multi-Agent Reinforcement Learning

Multi-agent reinforcement learning (MARL) encompasses a large body of literature extending RL techniques to multi-agent settings. In general, MARL algorithms subscribe to either decentralized or centralized learning. With decentralized learning (or independent learning), each agent is responsible for updating its own policy (or Q-network) individually [64]. Though recent work has shown that decentralized learning is feasible in complex environments [20], it is often challenging, due to the non-stationary (and therefore unstable) nature of the learning problem [38, 46]. Centralized techniques stabilize multi-agent learning by allowing agents to share a joint Q-network during training, then act independently at test-time [27, 44]. Centralization has also proven useful for emergent communication, as agents can share gradients directly through their communication channels [26]. Despite these benefits, we prioritize decentralized learning because it more accurately represents the learning problem that humans and animals face in the real-world.

Recently, significant effort has been aimed at connecting ideas from the curriculum learning literature [4] to RL [55]. Such methods have derived curricula from virtually every aspect of the RL problem; including reward shaping [3, 52], modifying initial state distributions [25], and procedurally-generating sub-tasks [54, 59]. Some multi-agent curricula have been shown to lead to more generally-capable RL agents [21]. Further work has examined the environment's role in generating curricula, leading to new methods such as environment shaping [15] and unsupervised environment design [22]. Of particular relevance is the work of Co-Reyes et al. [15], which considers curricula that manipulate the dynamics of the RL environment to the benefit of learning agents. Our work combines a velocity-based variant of environment shaping with an additional curriculum for bootstrapping multi-agent learning with single-agent experience, which itself is inspired by the strategy employed by Yang et al. [71] for multi-goal learning.

2.3 Pursuit-Evasion

Pursuit-evasion is a classic game setting for studying multi-agent coordination [33]. Though often played on a graph [51], work on continuous-space pursuit-evasion has enabled real-world applications such as unmanned aerial vehicles [68] and mobile robots [14]. Further work has shown that optimal control strategies can be derived from value functions [34] or even learned from scratch in MARL setting [44]. A relevant class of pursuit-evasion games define the evader to be of equal or greater speed than the pursuers. This setting highlights the need for coordinated motion (e.g. encircling) [67] and communication [69] by the pursuers and is subject to theoretical performance bounds under these conditions [57]. Our work uses this setting to study the emergence of animalistic coordination strategies and is similar to Janosov et al. [35] in that regard, but with an emphasis on learned, rather than hand-defined, strategies.

3 PRELIMINARIES

3.1 Markov Game

A Markov game extends the Markov Decision Process (MDP) to multi-agent settings [42]. For n agents, the formalism consists of a joint action-space $\mathcal{A} = \{\mathcal{A}_1, ..., \mathcal{A}_n\}$ and joint observation space $O = \{O_1, ..., O_n\}$ that define the actions and observations, respectively. The environment defines a state space $\mathcal{S}$ from which a state s_t is drawn each time-step. The environment transitions from its current state s_t to a new state s_{t+1} when actions are selected by each of the agents, as dictated by the transition function $T : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$. Finally, a reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ scores the quality of the composite action of the agents. The goal of each agent is to maximize the total reward it receives over time.

3.2 Deep Deterministic Policy Gradients

Deep Deterministic Policy Gradients (DDPG) is an off-policy actor-critic algorithm for policy gradient learning in continuous action spaces [41]. DDPG learns an optimal deterministic policy μ_ϕ with respect to the RL objective:

$$J(\phi) = \mathbb{E}_s[Q_\omega(s, a)|_{s=s_t, a=\mu_\phi(s_t)}] \tag{1}$$

by performing gradient ascent over the following gradient:

$$\nabla_\phi J(\phi) = \mathbb{E}_s[\nabla_a Q_\omega(s, a)|_{s=s_t, a=\mu(s_t)} \nabla_\phi \mu(s)|_{s=s_t}] \tag{2}$$

where ϕ and ω are parameters associated with μ and Q, respectively. Equation (2) is a consequence of the deterministic policy gradient theorem [63]. For critic updates, DDPG minimizes the loss function:

$$L(\omega) = \mathbb{E}_{s,a,r,s'}[(Q_\omega(s, a) - (r(s, a) + \gamma Q_\omega(s', \mu_\phi(s'))))^2] \tag{3}$$

where (s, a, r, s') are transition tuples sampled from a replay buffer and $\gamma \in [0, 1]$ is a scalar discount factor. In this work, agents learn in a decentralized manner, each performing DDPG updates individually.

3.3 Potential Field Navigation

Given an agent i with position q_i, we can define a potential function $U(q_i, q_{\text{goal}})$ between q_i and a target point q_{goal} such that the negative gradient $F(q_i, q_{\text{goal}}) = -\nabla U(q_i, q_{\text{goal}})$ specifies a control law for i's motion. For example, let $U_{\text{att}}(q_i, q_{\text{goal}})$ be a quadratic function of distance between q_i and a target point q_{goal}:

$$U_{\text{att}}(q_i, q_{\text{goal}}) = \frac{1}{2}k_{\text{att}} d(q_i, q_{\text{goal}})^2 \tag{4}$$

where k_{att} is an attraction coefficient and $d(,)$ is a measure of distance. The resulting force exerted on agent i is:

$$F_{\text{att}} = -\nabla U_{\text{att}}(q_i, q_{\text{goal}}) = -k_{\text{att}}(q_i - q_{\text{goal}}) \tag{5}$$

In this work, the agent's action-space is defined in terms of headings, so only the *direction* of this force impacts our agents.

3.4 Pursuit-evasion

Pursuit-evasion (i.e. predator-prey) is a classic setting for studying multi-agent coordination [33]. A pursuit-evasion game is defined between n pursuers $\{p_1, ..., p_n\}$ and a single evader e. The goal of the pursuers is to catch the evader as quickly as possible and, conversely, the goal of the evader is to remain uncaught. Each agent i is described by its current position and heading q_i and is subject to planar motion $\dot{q}_i$:

$$q_i = \begin{bmatrix} x_i \\ y_i \\ \theta_i \end{bmatrix} \qquad \dot{q}_i = \begin{bmatrix} \dot{x}_i \\ \dot{y}_i \\ \dot{\theta}_i \end{bmatrix} = \begin{bmatrix} |\vec{v}_i| \cos(\theta_i) \\ |\vec{v}_i| \sin(\theta_i) \\ \text{atan2}(\dot{y}_i, \dot{x}_i) \end{bmatrix}$$

where $\vec{v}_i$ is the agent i's velocity. The environment state s_t is described by the position and heading of all agents $s_t = \{q_{p_1}, ..., q_{p_n}, q_e\}$. Upon observing s_t, each agent selects its next heading θ_i as an action. The chosen heading is pursued at the maximum allowed speed for each agent ($|\vec{v}_p|$ for the pursuers, $|\vec{v}_e|$ for the evader); with orientation changes being instantaneous. To encourage teamwork, we set $|\vec{v}_p| \le |\vec{v}_e|$. We assume the evader to be part of the environment, as defined by the potential function:

$$U_{\text{evade}}(\theta_e) = \sum_i \left(\frac{1}{r_i}\right) \cos(\theta_e - \tilde{\theta}_i) \tag{6}$$

where r_i and $\tilde{\theta}_i$ are the L2-distance and relative angle between the evader and the i-th pursuer, respectively, and θ_e is the heading of the evader. This objective is inspired by theoretical analysis of escape strategies in the pursuit-evasion literature [57]. Intuitively, Equation (6) pushes the evader away from pursuers, taking the largest bisector between any two when possible. The goal of the pursuers—to capture the evader as quickly as possible—is mirrored in the reward function, where $r(s_t, a_t) = 50.0$ if the evader is captured and $r(s_t, a_t) = -0.1$ otherwise. A derivation of Equation (6) and additional environmental details are provided in Appendix A.

3.5 Implicit Communication

For clarity, we restate the definitions of implicit vs. explicit communication from Breazeal et al. [7] and implicit signals from De Greeff and Nolfi [19], which we adopt throughout this work.

DEFINITION 3.1 (EXPLICIT COMMUNICATION). *Communication that is "deliberate where the sender has the goal of sharing specific information with the collocutor" [7].*

DEFINITION 3.2 (IMPLICIT COMMUNICATION). *Communication that conveys "information that [is] inherent in behavior but which is not deliberately communicated" [7].*

Definition 3.3 (Implicit Signal). *The "signal that is generated by the actual physical position of the [agents] and that is detected by the other [agents]" [19].*

In the context of RL agents, an implicit signal refers to positional information observed by an agent's teammates as part of the environmental state s_t or observation o_t space, whereas explicit communication involves sending/receiving messages over a dedicated communication channel. We note that Definitions 3.1 and 3.2 rely on deliberation (or intention), whereas Definition 3.3 does not. For the purposes of our study—exploring the first-step in a bottom-up approach to emergent communication—we therefore focus on implicit signals as a form of emergent implicit communication.

3.6 Instantaneous Coordination

Instantaneous Coordination (IC) is a measure of social influence between agents [36]. IC is defined for two agents i and j as the mutual information $\text{I}(a_i^t; a_j^{t+1})$ between i's action at time t and j's action at the next time-step. Formally, assuming agent i's actions are drawn from the random variable A_i with marginal distribution P_{A_i} (and similarly for agent j), we can rewrite IC using the standard definition of mutual information as the Kullback-Leibler divergence between the joint distribution and the product of the marginals:

$$I(A_i; A_j) = D_{\text{KL}}(P_{A_i A_j} || P_{A_i} \times P_{A_j})$$
$$= \sum_{\substack{a_i \in \mathcal{A}_i, \\ a_j \in \mathcal{A}_j}} P_{A_i A_j}(a_i, a_j) \log \left(\frac{P_{A_i A_j}(a_i, a_j)}{P_{A_i}(a_i) \times P_{A_j}(a_j)} \right)$$

where $\mathcal{A}_i$ and $\mathcal{A}_j$ are the spaces over A_i and A_j, respectively. Intuitively, high IC is indicative of influential behavior, while low IC indicates that agents are acting independently. We highlight that, in the absence of explicit communicative actions, IC is a measure of implicit signals (as defined in the previous section). This is because agent i's action at time t inherently dictates i's position q_i at time $t + 1$, which is observed by i's teammates. IC in this context is therefore a measure of position-based social influence. Since this work considers deterministic policies and uses IC at test-time, IC is computed over Monte-Carlo estimates of the relevant distributions.

4 METHOD

Our goal is to learn multi-agent coordination with decentralized training and, in doing so, explore the role of implicit communication in teamwork. However, there are a number of issues with decentralized learning in difficult, sparse reward environments, such as our pursuit-evasion game. First, though setting $|\vec{v}_p| \leq |\vec{v}_e|$ is important for studying teamwork, it places the pursuers at a severe deficit. In the early stages of training—since each pursuer's action selection is determined by the randomly initialized weights of its policy network—the chance of slower pursuers capturing the evader defined in Equation (6) is extremely low. The pursuers are unlikely to obtain a positive reward signal, which is vital to improving their policies. This issue is exacerbated by non-stationarity. In the case of decentralized DDPG, multiple agents learning in the same environment causes the value of state-action pairs for any one agent (as judged by its Q-function) to change as a result of policy updates of *other* agents. This non-stationarity leads to higher-variance gradient estimates and unstable learning. Though recent advances in

"centralized training, decentralized execution" help in such cases [44], they violate our goal of decentralized learning.

To address these challenges, we introduce a *curriculum-driven method* for decentralized multi-agent learning in sparse reward environments. Our strategy is motivated by the following principles: (i) we can create a *curriculum over task difficulty by shaping the environment* to the current skill level of the cooperative agents; (ii) we can aide agents in learning multi-agent team behaviors by forcing them to *learn single-agent behaviors first*. More specifically, we combine an environment shaping curriculum over agent velocities with a behavioral curriculum for bootstrapping cooperative multi-agent learning with successful single-agent experience. We refer to this curriculum-driven variant of DDPG as CD-DDPG throughout the rest of this paper.

4.1 Velocity Ratio Curriculum

Curriculum learning [4] is a popular technique for solving complex learning problems by breaking them down into smaller, easier to accomplish tasks. Recent advances in ecological reinforcement learning [15] have introduced environment shaping—in which properties, initial states, or the dynamics of an environment are modified gradually—as a useful and more natural instantiation of curriculum learning for RL agents (than, say, traditional reward shaping). With this in mind, we construct a sequence of increasingly difficult pursuit-evasion environments by incrementally lowering the ratio of pursuer speed to evader speed ($|\vec{v}_p|/|\vec{v}_e|$).

More formally, we define a curriculum over velocity bounds. Let $\vec{v}_0$ be an initial setting of the environment's velocity ratio $|\vec{v}_p|/|\vec{v}_e|$. We anneal $|\vec{v}_p|/|\vec{v}_e|$ to a target ratio $\vec{v}_{\text{target}}$ over v_{decay} epochs as:

$$\vec{v}_i \leftarrow \vec{v}_{\text{target}} + (\vec{v}_0 - \vec{v}_{\text{target}}) * \max \left(\frac{(v_{\text{decay}} - i)}{v_{\text{decay}}}, 0.0 \right) \quad (7)$$

where i represents the current training epoch and, in turn, $\vec{v}_i$ the current velocity ratio. In practice, we initialize the environment such that $\vec{v}_0 > 1.0$ (i.e. pursuers are faster than the evader), then anneal this ratio slowly as training progresses. This gives the pursuers an opportunity enjoy a velocity advantage early on in training, then develop increasingly coordinated strategies to capture the evader as $|\vec{v}_p|/|\vec{v}_e|$ decays.

4.2 Behavioral Curriculum

We combine the aforementioned velocity curriculum with a behavioral curriculum that extends off-policy learning to allow for targeted single-agent exploration early in the training process. As a consequence of the deterministic policy gradient and importance sampling [63], the policy gradient can be estimated in an off-policy manner—i.e. using trajectories sampled from a separate behavior policy $\beta(a|s)$ where $\beta(a|s) \neq \mu_\phi$. Formally, this means Equation (2) can be represented equivalently as:

$$J_\beta(\phi) = \int_{\mathcal{S}} p^\beta Q_\omega(s, a)|_{s=s_t, a=\mu_\phi(s_t)} \, ds$$
$$= \mathbb{E}_{s \sim p^\beta} [Q_\omega(s, a)|_{s=s_t, a=\mu_\phi(s_t)}]$$

with the corresponding gradient:

$$\nabla_\phi J_\beta(\phi) = \mathbb{E}_{s \sim p^\beta} [\nabla_a Q_\omega(s, a)|_{s=s_t, a=\mu(s_t)} \nabla_\phi \mu(s)|_{s=s_t}]$$

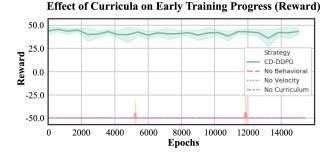**Effect of Curricula on Early Training Progress (Reward)**

Figure 2: An ablation study of curriculum-driven learning early in the training process. Note that $|\vec{v}_p|/|\vec{v}_e|$ is decreasing over time (increasing the task difficultly), so learning is characterized by *sustained* reward rather than *improved reward*. Only CD-DDPG successfully learns team coordination.

where p^β is the state distribution of the behavior policy β.

Our behavioral curriculum takes advantage of off-policy learning by splitting training into two exploration phases that use distinct behavior policies β_0 and β_μ, respectively. The key is that we define β_0 strategically to be a supervisory policy that collects *successful single-agent experience* and β_μ to be a standard exploration policy (ϵ−greedy for discrete actions, random noise for continuous actions). In multi-agent settings, defining β_0 this way allows each agent to learn how to interact with the environment first, before learning to coordinate with teammates. Without this curriculum, agents are forced to learn both simultaneously.

In our pursuit-evasion environment, we define β_0 as follows:

$$\beta_0 = -\nabla U_{\text{att}}(q_{p_i}, q_e) \qquad (8)$$

where U_{att} is the attractive potential-field defined in Equation (4). Note that β_0 is a greedy policy that runs directly towards the evader. This strategy is obviously sub-optimal when $|\vec{v}_p|/|\vec{v}_e| \leq 1.0$, but helps pursuers learn to move in the direction of the evader when $|\vec{v}_p|/|\vec{v}_e| > 1.0$. In our experiments, we use β_0 only during the first phase of the velocity curriculum. After this phase, agents follow the standard DDPG behavior policy $\beta_\mu = \mu_\phi(s_t) + \mathcal{N}$, where $\mathcal{N}$ is the Ornstein-Uhlenbeck noise process.

5 RESULTS

Our evaluation addresses two primary questions: (i) Does our curriculum-driven strategy enable decentralized agents to learn to coordinate in difficult sparse reward environments? (ii) To what extent does implicit communication emerge in the learned strategy? To answer these questions, we first perform an ablation study over each of the multi-agent curricula that comprise CD-DDPG. Then, we measure the performance of CD-DDPG against a set of analytical and learned strategies of increasing sophistication. We intentionally select these strategies to represent ablations over both coordination and implicit communication.

5.1 Multi-Agent Curricula

To isolate our method's performance, we run ablations over each of CD-DDPG's constituent parts. We compare CD-DDPG to the following alternatives: *No Behavioral*, which follows the velocity

Table 1: A summary of the strategies used to evaluate CD-DDPG and their capacity for coordination and implicit communication. Outperforming one of these methods is equivalent to surpassing that method's level of sophistication in coordination, capacity for implicit communication, or both.

Name	Coordination	Implicit Communication
Greedy	No	No
CD-DDPG (Partial)	Yes	No
Pincer	Yes	Yes

curriculum, but not the behavioral one; *No Velocity*, which follows the behavioral curriculum, but not the velocity one; and *No Curriculum*, which uses neither curriculum during training (i.e. it is vanilla DDPG trained at a constant ratio $|\vec{v}_p|/|\vec{v}_e| = 0.7$). We trained each method for 15,000 epochs in our pursuit-evasion environment across 10 different random seeds. Results are shown in Figure 2.

No Curriculum, unsurprisingly, flat-lines throughout the training process. It never experiences a positive reward signal, reflecting the difficulty of pursuit-evasion when $|\vec{v}_p|/|\vec{v}_e| < 1.0$. *No Behavioral*, trained from a velocity ratio of $|\vec{v}_p|/|\vec{v}_e| = 1.2$ downwards, experiences small spikes in reward, but not consistently enough to improve the pursuers' policies. Finally, *No Velocity* also fails to gain any traction (it's curve is behind No Curriculum in the figure), highlighting the importance of both curricula working together. This validates our intuition from Section 4. Even with a velocity advantage, the agents struggle to learn individual interaction with the environment and coordination with their teammates simultaneously. *CD-DDPG* is able to capture valuable experience even in the earliest stages of training. This bootstraps each pursuer's learning process, allowing them to maintain a high level of performance even after the warm-up is over. We therefore find that the combination of both curricula is crucial to warm-starting policy learning.

5.2 Properties of Emergent Teamwork

To study the strength of CD-DDPG's coordination further and identify the role of implicit communication in the agents' emergent teamwork, we compare the performance of CD-DDPG to a set of analytical and learned strategies of increasing sophistication. These strategies represent ablations over both coordination and implicit communication—i.e. outperforming one of these methods is equivalent to surpassing that method's level of sophistication in coordination, capacity for implicit communication, or both. With this in mind, we evaluate CD-DDPG against the following policies, which are also summarized in Table 1[1]:

Greedy. Each pursuer follows the greedy control strategy in Equation (4). In Greedy pursuit, each pursuer ignores the positions of its teammates. Greedy pursuit therefore represents independent action (i.e. no coordination, no communication).

CD-DDPG (Partial). We train CD-DDPG under partial observability. Instead of the complete environment state s_t, each pursuer p_i receives a private observation $o_t = \{q_{p_i}, q_e\}$ consisting of its

---

[1]We do not evaluate against centralized methods such as MADDPG [44] because they violate our requirement of decentralized learning.
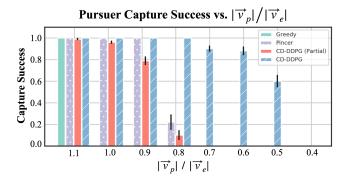
Figure 3: Capture success as a function of velocity. CD-DDPG succeeds at $|\vec{v}_p|/|\vec{v}_e| = 0.5$ whereas each competing method fails below $|\vec{v}_p|/|\vec{v}_e| = 0.8$.

own location q_{p_i} and the location of the evader q_e. Despite not observing each other, CD-DDPG (Partial) pursuers are capable of coordinating through static role assignment. This is equivalent to assigning roles before each trajectory—i.e. p_1 always approaches from the left, p_2 from the right, etc.—and coordinating through these roles during pursuit. CD-DDPG (Partial) pursuers are therefore coordinated, but with no ability to communicate implicitly to modify their behavior extemporaneously.

Pincer. We define the Pincer strategy as an adversarial function that exploits knowledge of the evader's objective in Equation (6):

$$F(\tilde{\boldsymbol{\theta}_i}, \boldsymbol{r_i}) = \max_{\tilde{\boldsymbol{\theta}_i}, \boldsymbol{r_i}} \Big[\min_{\theta_e} \big[U_{\text{evade}}(\theta_e) \big] \Big] \qquad (9)$$

$$= \max_{\tilde{\boldsymbol{\theta}_i}, \boldsymbol{r_i}} \Big[\min_{\theta_e} \Big[\sum_i \Big(\frac{1}{r_i} \Big) \cos(\theta_e - \tilde{\theta}_i) \Big] \Big] \qquad (10)$$

where $\tilde{\boldsymbol{\theta}_i}$ and $\boldsymbol{r_i}$ are the polar coordinates of each pursuer relative the evader. Intuitively, Pincer encircles the evader and cuts off potential bisector escape paths while enclosing the circle. It therefore supports both coordination—pursuers uphold a circular formation—and implicit communication—every time-step, pursuers use information from the locations of their teammates to adjust their own position on the circular formation. We provide additional details on the Pincer strategy in Appendix A.3.

All experiments involve $n = 3$ pursers. For CD-DDPG and CD-DDPG (Partial), agents are trained for $50,000$ epochs. The velocity ratio is decreased from $|\vec{v}_p|/|\vec{v}_e| = 1.2$ to $|\vec{v}_p|/|\vec{v}_e| = 0.4$ in decrements of 0.1 (i.e. eight separate training sessions each). Velocity ratio decay occurs over $v_{\text{decay}} = 15000$ epochs, during which the behavioral curriculum β_0 is used (but only during the first training session). After training, we test the resulting policies at each velocity step. Test-time performance is measured over 100 independent trajectories (averaged over five random seeds each).

5.2.1 Capture Success. First, we evaluate capture success as a function of the velocity advantage of the evader. The results are shown in Figure 3. Unsurprisingly, each method has a high capture success rate when $|\vec{v}_p|/|\vec{v}_e| > 1.0$. The Greedy strategy drops off at $|\vec{v}_p|/|\vec{v}_e| = 1.0$, which is also expected—a straight-line chase only works when $|\vec{v}_p|/|\vec{v}_e| > 1.0$. The Pincer and CD-DDPG (Partial)
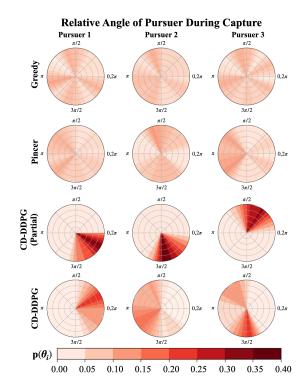

Figure 4: Distribution of relative angle for each pursuer during capture, binned in the range $[0, 2\pi]$ and displayed as a heatmap. Each row represents a pursuit strategy.

strategies are able to coordinate successfully at lower speeds, but eventually **fail to capture the evader at $|\vec{v}_p|/|\vec{v}_e| = 0.8$** and below. As $|\vec{v}_p|/|\vec{v}_e|$ decreases further, CD-DDPG significantly outperforms the other strategies. Specifically, CD-DDPG successfully completes the pursuit-evasion task at a **speed ratio of 0.5** (i.e. pursuers moving at half the evader's speed). These results show that CD-DDPG learns to coordinate significantly more effectively than the other strategies.

5.2.2 Relative position during capture. Next, we study the nature of each strategy's coordination with finer granularity. In sparse reward tasks, the most important time-steps are those that immediately precede reward. For this reason, we look at the distribution of pursuer locations at the time of capture. We collect 100 successful trajectories from each strategy and compute the distribution of pursuer positions relative the evader. In general, we expect coordinated pursuit to exhibit rotational symmetry during capture. Rotational symmetry suggests that pursuers have learned strategies which lead them to well-defined "capture points" around the evader. Conversely, rotational invariance is indicative of independent pursuit—i.e. pursuers do not follow concrete patterns of attack. Results for this study are shown in Figure 4.

We find that Greedy and Pincer both yield uniform capture distributions. This is unsurprising for Greedy pursuers, whose pursuit paths are not effected by their teammates. The Pincer strategy encircles the evader, but does not constrain pursuers to specific locations on the circle. This leads to less-structured capture(i.e. weaker role

Instantaneous Coordination vs. $|\overrightarrow{v}_p|/|\overrightarrow{v}_e|$
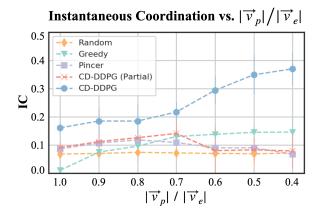

Figure 5: Instantaneous Coordination (IC) results as a function of velocity ratios. Agents trained with CD-DDPG exchange up to 0.375 bits of information per time-step on average, increasing as $|\vec{v}_p|/|\vec{v}_e|$ drops, whereas competing strategies peaks remains constant.

assignment) by the Pincer pursuers. In contrast, CD-DDPG (Partial) pursuers demonstrate very strong role assignment, with each pursuer capturing the evader from the same relative angle each time. Taken into context with the results from Figure 3, it is clear that this level of rotational symmetry impacts success. In fact, it is an example of over-commitment to role assignment. The pursuers adopt very constrained roles—e.g. "p_1 always move left", "p_2 always move right"—which works when $|\vec{v}_p|/|\vec{v}_e| \geq 1.0$, but fails at lower velocities. CD-DDPG balances rotational symmetry and invariance. Each pursuer follows a unique angle towards the evader, but does not commit to that angle completely. CD-DDPG therefore learns structured coordination while allowing pursuers to make dynamic adjustments to their position relative the evader to achieve capture.

5.2.3 Position-based Social influence. To further study the role of implicit communication in pursuer performance, we compute the IC score for each strategy. As noted in, section 3, by measuring the amount that one agent's actions (and therefore its next position) influences the actions of its teammates, IC quantifies the exchange of implicit signals (see Definition 3.3) amongst teammates. Following Jaques et al. [36], we compute IC empirically as a Monte-Carlo approximation over multi-agent trajectories. We average influence across all trajectory steps and for each agent-agent pair. We also evaluate pursuers that act randomly, which provides a baseline for independent action. The results are shown in Figure 5.

We find that, as $|\vec{v}_p|/|\vec{v}_e|$ decreases, the IC levels attained by CD-DDPG increases significantly, whereas it remains stagnant for other methods. In fact, across all $|\vec{v}_p|/|\vec{v}_e|$ levels, we find that pursuers trained with **CD-DDPG exchange up to 0.375 bits of information** per time-step on average, compared to a maximum of **only 0.15 bits on average** from the baseline methods. This indicates that CD-DDPG *achieves increasingly complex coordination* as task difficulty increases and is a promising sign that the CD-DDPG team is exchanging implicit signals—i.e. each pursuer is responding to positional information from its teammates. Finally, we note a minor (though surprising) increase in coordination for the greedy pursuers

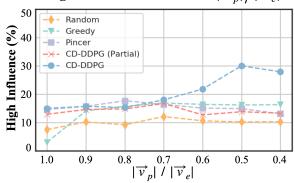## High Influence Moments vs. $|\overrightarrow{v}_p|/|\overrightarrow{v}_e|$

Figure 6: Percentage of high-influence moments as a function of velocity ratios. Agents trained with CD-DDPG become more influential as $|\vec{v}_p|/|\vec{v}_e|$ decreases.

at low velocities. This is an artifact we dub "phantom coordination" and discuss in the Appendix B.2.

5.2.4 High-influence moments. We also report on the percentage of high-influence moments that occur between pairs of agents (see Figure 6). A high-influence moment is a time-step in which IC is above the mean IC for the entire trajectory. Similar to previous work [36], we find that influence is sparse in most cases—**only between 10-15%** of trajectory steps exhibit high-influence across all $|\vec{v}_p|/|\vec{v}_e|$ levels. The exception, notably, occurs for CD-DDPG. At low speeds (i.e. $|\vec{v}_p|/|\vec{v}_e| \leq 0.7$), we see a significant increase in the percentage of high-influence moments between the CD-DDPG pursuers, reaching a **maximum near 30%**. This is further evidence that, as $|\vec{v}_p|/|\vec{v}_e|$ decreases, CD-DDPG pursuers form increasingly highly-coordinated formations and make split-second decisions based on the movements of their teammates. This points more concretely to the use of implicit signals between CD-DDPG pursuers.

Interestingly, the behavior of CD-DDPG closely matches the documented behaviors of social predators such as dolphins and wolves—i.e. sudden changes of position/orientation as a response to the movements of other teammates [32]. We elaborate on these and other qualitative findings in Appendix B.

6 CONCLUSION

This work connects emergent communication to the spectrum of communication that exists in nature, highlighting the importance of interpreting communication as a spectrum from implicit to explicit communication. We proposed a curriculum-driven strategy for policy learning in difficult multi-agent environments. Experimentally, we showed that our curriculum-driven strategy enables pursuers to coordinate and capture a superior evader, outperforming other highly-sophisticated analytic pursuit strategies. We also provided evidence suggesting that the emergence of implicit communication is a key contributor to the success of this strategy. There are a number of extensions of this work that study how common principles contribute to integrated, communicative behavior; including: imperfect state information, increased environmental complexity, and nuanced social dynamics between agents.

REFERENCES

[1] B Baker, I Kanitscheider, T Markov, et al. 2019. Emergent tool use from multi-agent autocurricula. *arXiv preprint:1909.07528* (2019).

[2] N Bard, J N Foerster, et al. 2020. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence* (2020).

[3] M Bellemare, S Srinivasan, G Ostrovski, T Schaul, D Saxton, and R Munos. 2016. Unifying count-based exploration and intrinsic motivation. *NeurIPS* (2016).

[4] Y Bengio, J Louradour, R Collobert, et al. 2009. Curriculum learning. In *ICML*.

[5] C Boesch and H Boesch. 1989. Hunting behavior of wild chimpanzees in the Tai National Park. *AJPA* (1989).

[6] J Bradbury, S Vehrencamp, et al. 1998. Principles of animal communication. (1998).

[7] C Breazeal, C Kidd, A Thomaz, G Hoffman, and M Berlin. 2005. Effects of non-verbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ*. IEEE, 708–713.

[8] R Bshary, A Hohner, K Ait-el Djoudi, et al. 2006. Interspecific communicative and coordinated hunting between groupers and giant moray eels in the Red Sea. *PLoS Biol* (2006).

[9] A Butchibabu, C Sparano-Huiban, L Sonenberg, and J Shah. 2016. Implicit coordination strategies for effective team communication. *Human factors* 58, 4 (2016), 595–610.

[10] R Chaabouni, E Kharitonov, D Bouchacourt, et al. 2020. Compositionality and generalization in emergent languages. *arXiv preprint:2004.09124* (2020).

[11] R Chaabouni, E Kharitonov, E Dupoux, et al. 2019. Anti-efficient encoding in emergent communication. In *NeurIPS*.

[12] Y Che, A Okamura, and D Sadigh. 2020. Efficient and trustworthy social navigation via explicit and implicit robot–human communication. *IEEE Transactions on Robotics* 36, 3 (2020), 692–707.

[13] N Choi, J Kim, N Kokubun, et al. 2017. Group association and vocal behaviour during foraging trips in Gentoo penguins. *Scientific Reports* (2017).

[14] T Chung, G Hollinger, and V Isler. 2011. Search and pursuit-evasion in mobile robotics. *Autonomous robots* (2011).

[15] J Co-Reyes, S Sanjeev, G Berseth, A Gupta, and S Levine. 2020. Ecological Reinforcement Learning. *arXiv preprint arXiv:2006.12478* (2020).

[16] P Cohen and H J Levesque. 1991. Teamwork. *Nous* 25, 4 (1991), 487–512.

[17] A Cowen-Rivers and J Naradowsky. 2020. Emergent Communication with World Models. *arXiv preprint:2002.09604* (2020).

[18] A Das, T Gervet, J Romoff, et al. 2019. Tarmac: Targeted multi-agent communication. In *ICML*.

[19] J De Greeff and S Nolfi. 2010. Evolution of implicit and explicit communication in mobile robots. In *Evolution of communication and language in embodied agents*. Springer, 179–214.

[20] C de Witt, T Gupta, D Makoviichuk, V Makoviychuk, P Torr, M Sun, and S Whiteson. 2020. Is Independent Learning All You Need in the StarCraft Multi-Agent Challenge? *arXiv preprint arXiv:2011.09533* (2020).

[21] Open Ended Learning Team DeepMind. 2021. Open-ended learning leads to generally capable agents. *arXiv preprint arXiv:2107.12808* (2021).

[22] M Dennis, N Jaques, E Vinitsky, et al. 2020. Emergent complexity and zero-shot transfer via unsupervised environment design. *NeurIPS* (2020).

[23] T Eccles, Y Bachrach, G Lever, et al. 2019. Biases for emergent communication in multi-agent reinforcement learning. In *NeurIPS*.

[24] E Entin and D Serfaty. 1999. Adaptive team coordination. *Human factors* (1999).

[25] C Florensa, D Held, M Wulfmeier, M Zhang, and P Abbeel. 2017. Reverse curriculum generation for reinforcement learning. In *CoRL*, 482–495.

[26] J Foerster, I Assael, N De Freitas, et al. 2016. Learning to communicate with deep multi-agent reinforcement learning. In *NeurIPS*.

[27] J Foerster, G Farquhar, T Afouras, N Nardelli, and S Whiteson. 2018. Counterfactual multi-agent policy gradients. In *AAAI*, Vol. 32.

[28] J Foerster, F Song, E Hughes, et al. 2018. Bayesian action decoding for deep multi-agent reinforcement learning. *arXiv preprint:1811.01458* (2018).

[29] N Gildert, A Millard, A Pomfret, and J Timmis. 2018. The need for combining implicit and explicit communication in cooperative robotic systems. *Frontiers in Robotics and AI* 5 (2018), 65.

[30] P Grover and A Sahai. 2010. Implicit and explicit communication in decentralized control. In *Allerton Conference on Communication, Control, and Computing*. IEEE, 278–285.

[31] S Havrylov and I Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *NeurIPS*.

[32] J Herbert-Read. 2016. Understanding how animal groups achieve coordinated movement. *JEB* (2016).

[33] R Isaacs. 1999. *Differential games: a mathematical theory with applications to warfare and pursuit, control and optimization*. Courier Corporation.

[34] J Jang and C Tomlin. 2005. Control strategies in multi-player pursuit and evasion game. In *AIAA ARC*.

[35] M Janosov, C Virágh, G Vásárhelyi, et al. 2017. Group chasing tactics: how to catch a faster prey. *NJP* (2017).

[36] N Jaques, A Lazaridou, E Hughes, et al. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *ICML*.

[37] R Knepper, C Mavrogiannis, J Proft, and C Liang. 2017. Implicit communication in a joint action. In *HRI*. 283–292.

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

[39] A Lazaridou and M Baroni. 2020. Emergent Multi-Agent Communication in the Deep Learning Era. *arXiv preprint:2006.02419* (2020).

[40] C Liang, J Proft, E Andersen, and R Knepper. 2019. Implicit communication of actionable information in human-ai teams. In *CHI*.

[41] T Lillicrap, J Hunt, A Pritzel, et al. 2015. Continuous control with deep reinforcement learning. *arXiv preprint:1509.02971* (2015).

[42] M Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In *ICML*.

[43] R Lowe, J Foerster, et al. 2019. On the pitfalls of measuring emergent communication. *arXiv preprint:1903.05168* (2019).

[44] R Lowe, Yi I Wu, A Tamar, et al. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *NeurIPS*.

[45] W Mason and J Hollis. 1962. Communication between young rhesus monkeys. *Animal Behaviour* (1962).

[46] L Matignon, G J Laurent, and N Le Fort-Piat. 2012. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *The Knowledge Engineering Review* (2012).

[47] I Mordatch and P Abbeel. 2017. Emergence of grounded compositional language in multi-agent populations. *arXiv preprint:1703.04908* (2017).

[48] C Muro, R Escobedo, L Spector, et al. 2011. Wolf-pack hunting strategies emerge from simple rules in computational simulations. *Behavioural processes* (2011).

[49] D Neumann and M Orams. 2003. Feeding behaviours of short-beaked common dolphins, Delphinus delphis, in New Zealand. *Aquatic Mammals* (2003).

[50] E Pagello, A D'Angelo, et al. 1999. Cooperative behaviors in multi-robot systems through implicit communication. *Robotics and Autonomous Systems* (1999).

[51] T Parsons. 1978. Pursuit-evasion in a graph. In *TAG*.

[52] D Pathak, P Agrawal, A Efros, and T Darrell. 2017. Curiosity-driven exploration by self-supervised prediction. In *ICML*.

[53] R Peterson and P Ciucci. 2003. The wolf as carnivore. *Wolves: Behavior, Ecology, and Conservation* (2003).

[54] R Portelas, C Colas, K Hofmann, and P Oudeyer. 2020. Teacher algorithms for curriculum learning of deep rl in continuously parameterized environments. In *CoRL*. PMLR, 835–853.

[55] R Portelas, C Colas, L Weng, K Hofmann, and P Oudeyer. 2020. Automatic curriculum learning for deep rl: A short survey. *arXiv preprint arXiv:2003.04664* (2020).

[56] N Quick and V Janik. 2012. Bottlenose dolphins exchange signature whistles when meeting at sea. *Biological Sciences* (2012).

[57] M Ramana and M Kothari. 2017. Pursuit-evasion games of high speed evader. *JINT* (2017).

[58] C Resnick, A Gupta, J Foerster, et al. 2019. Capacity, bandwidth, and compositionality in emergent language learning. *arXiv preprint:1910.11424* (2019).

[59] S Risi and J Togelius. 2019. Procedural content generation: From automatically generating game levels to increasing generality in machine learning. (2019).

[60] E Sampaio, M Seco, R Rosa, et al. 2020. Octopuses punch fishes during collaborative interspecific hunting events. *Ecology* (2020).

[61] G Schaller. 2009. *The Serengeti lion: a study of predator-prey relations*. UChicago Press.

[62] A Schmidt. 2000. Implicit human computer interaction through context. *Personal technologies* (2000).

[63] D Silver, G Lever, N Heess, T Degris, D Wierstra, and M Riedmiller. 2014. Deterministic policy gradient algorithms.

[64] M Tan. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *ICML*.

[65] A Vail, A Manica, and R Bshary. 2013. Referential gestures in fish collaborative hunting. *Nature Communications* (2013).

[66] A Vaswani, N Shazeer, N Parmar, et al. 2017. Attention is all you need. In *NeurIPS*.

[67] T Vicsek. 2010. Closing in on evaders. *Nature* (2010).

[68] R Vidal, O Shakernia, H Kim, et al. 2002. Probabilistic pursuit-evasion games: theory, implementation, and experimental evaluation. *IEEE TRA* (2002).

[69] Y Wang, L Dong, and C Sun. 2020. Cooperative control for multi-player pursuit-evasion games with reinforcement learning. *Neurocomputing* (2020).

[70] J Wuertz, S Alharthi, W Hamilton, S Bateman, C Gutwin, A Tang, Z Toups, and J Hammer. 2018. A design framework for awareness cues in distributed multiplayer games. In *CHI*.

[71] J Yang, A Nakhaei, D Isele, K Fujimura, and H Zha. 2018. Cm3: Cooperative multi-goal multi-stage multi-agent reinforcement learning. *arXiv preprint arXiv:1809.05188* (2018).

[72] W Zhou, W Yan, and X Zhang. 2017. Collaboration for success in crowdsourced innovation projects: Knowledge creation, team diversity, and tacit coordination. In *HICSS*.