Curriculum-Driven Multi-Agent Learning and the Role of Implicit Communication in Teamwork

Niko A. Grupen\textsuperscript{1}  Daniel D. Lee\textsuperscript{1,2}  Bart Selman\textsuperscript{1}  \\
\textsuperscript{1}Cornell University  \textsuperscript{2}Cornell Tech  \\
niko@cs.cornell.edu

Abstract

We propose a curriculum-driven learning strategy for solving difficult multi-agent coordination tasks. Our method is inspired by a study of animal communication, which shows that two straightforward design features (mutual reward and decentralization) support a vast spectrum of communication protocols in nature. We highlight the importance of similarly interpreting emergent communication as a spectrum. We introduce a toroidal, continuous-space pursuit-evasion environment and show that naive decentralized learning does not perform well. We then propose a novel curriculum-driven strategy for multi-agent learning. Experiments with pursuit-evasion show that our approach enables decentralized pursuers to learn to coordinate and capture a superior evader, significantly outperforming sophisticated analytical policies. We argue through additional quantitative analysis—including influence-based measures such as Instantaneous Coordination—that emergent implicit communication plays a large role in enabling superior levels of coordination.

1 Introduction

Communication is the cardinal tool of coordination. Communication enables humans and animals alike to coordinate on complex tasks, plan future sequences of tasks, negotiate the allocation of resources, and explain their actions. For this reason, understanding and recreating the process through which communication emerges has long been a goal of philosophy, linguistics, cognitive science, and AI. Recently, the connection between coordination and mutual reward has propelled the adoption of multi-agent reinforcement learning (MARL) as a computational framework for studying communication. Emergent communication in MARL is now a well-developed research area that examines the representations and social conditions necessary for communication to emerge in multi-agent populations. Many existing approaches target emergent protocols that mirror the structure and sophistication of human language [31]. It has been shown that, given additional learning biases, it is possible for agents to converge upon a communication protocol that exhibits language-like properties like compositionality [10, 46] and Zipf’s Law [9]. Though natural language is an appropriate target—achieving emergent natural language is equivalent to an AI system achieving the uniquely human ability to represent a combinatorial world using relevant abstractions—it is only one of a multitude of communication protocols that arise in nature.

Communication is useful when decentralized agents must coordinate to maximize mutual reward. In nature, evolutionary processes over these conditions have given rise to a diverse scope of inter-species communication protocols and occasionally cause communication to bridge species boundaries. For example, in the context of foraging, the grouper fish (an open-water hunter) is known to communicate with the moray eel (a reef hunter) through a variety of head and body shakes, inciting the eel to drive prey into the open from hidden crevices in the reef [7].

Natural communication is shaped by the coordination task at hand, environmental complexity, and the cognitive, perceptual, and behavioral capabilities of the agents involved. Before we can endow multi-
agent systems with language-like communication, we must understand the full scope of protocols that arise from these pressures. Towards this end, we present a brief overview of communication protocols that exist in the animal kingdom; specifically within the context of social foraging. We highlight the importance of interpreting communication as a complete spectrum and posit that modeling implicit communication is an important first step in emergent communication research.

We then study multi-agent coordination and the role of implicit communication through experiments with pursuit-evasion games [26]. To accurately model the conditions under which coordination and communication emerges in the animal kingdom, we prioritize decentralized learning. Centralized training, however, reduces the negative effects of non-stationarity in MARL settings [18]; so we must compensate in other ways. In addition to non-stationarity, there are challenges with multi-agent learning in sparse-reward environments that require coordinated teamwork. 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. Such challenges exist in our pursuit-evasion environment, as a sophisticated evader can easily avoid the random movements of the pursuers, preventing them from receiving a positive reward signal during training.

We address these issues by introducing a curriculum-driven approach to MARL. First, we adopt a curriculum learning strategy 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 Experience Warm-Starting (EWS), a method for warm-starting policy learning by seeding each agent’s replay buffer with valuable experience from a strategic behavior policy. We show that this curriculum-driven strategy enables decentralized MARL algorithms to solve difficult pursuit-evasion games and outperform sophisticated analytic pursuit methods. Unlike other methods that incentivize communication through additional objectives [17] or rewards [29], our method learns using only environmental reward. Through additional quantitative analysis, we show evidence that implicit communication emerges as a key component of the agents’ coordination.

Preview of contributions Our work is summarized by the following contributions:

- We introduce a curriculum-driven learning strategy that enables cooperative agents to solve difficult coordination tasks. We apply our strategy to Deep Deterministic Policy Gradients (DDPG) [32] to create a new algorithm, which we call Curriculum-Driven DDPG.

- We introduce a novel toroidal, continuous space pursuit-evasion environment for studying multi-agent coordination.

- We show that our curriculum-driven strategy learns to solve difficult coordination tasks with sparse reward. Using our strategy, multiple slower-moving pursuers learn to coordinate to capture a superior evader, significantly outperforming sophisticated analytical methods.

- We perform additional quantitative analysis—including through influence-based measures like Instantaneous Coordination [29]—to study the role of implicit communication in the learned coordination strategies.
2 Preliminaries

Multi-agent reinforcement learning In multi-agent RL, a Markov game extends Markov Decision Processes (MDP) to the multi-agent setting [43]. The formalism consists of a joint action-space $\mathcal{A} = \{A_1, ..., A_N\}$ and joint observation space $\mathcal{O} = \{O_1, ..., O_N\}$ that define the actions and observations for $N$ agents, 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 $T : \mathcal{S} \times \mathcal{A}_1 \times ... \times \mathcal{A}_N \to \mathcal{S}$. Finally, a reward function $r : \mathcal{S} \times \mathcal{A}_1 \times ... \times \mathcal{A}_N \to \mathbb{R}$ 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. Given a policy $\pi_\phi$ with parameters $\phi$, this goal can be restated as finding the optimal parameters $\phi^*$ that maximize reward:

$$\phi^* = \arg \max_{\phi} J(\phi) = \arg \max_{\phi} \mathbb{E}_{\tau \sim p_\phi(r)} \left[ \sum_{t=0}^{T} \gamma^t r(s_t, a_t) \right]$$

where $a_t \sim \pi_\phi$ is the agent’s action, $r(s_t, a_t)$ is the reward, and $\gamma \in [0, 1]$ is a scalar discount factor.

Potential field navigation If we assume that each agent $i$ is a particle and the environment is represented by a potential field, we can define a potential function $U(q_i)$ such that the negative gradient $\nabla U(q_i) = -\nabla U(q_i)$ specifies a control law for agent $i$’s motion. For example, let $U_{\text{att}}(q_i)$ be a quadratic function of distance between $q_i$ and a target point $q_{\text{goal}}$:

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

(1)

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

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

(2)

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

Animal communication Natural language is the “grand-prize” for emergent communication, but it is only one example of a protocol that has emerged under evolutionary pressure. For example, in the context of foraging, weaver ant pheromone trails [2] and honeybee “waggle dances” [52] are well-known examples of communication. Further evidence has shown that grouper fin flares [7], lionfish body shakes [24], octopus punches [47], and a range of vocalizations (e.g. aplomado falcons [24], tai chimpanzees [5], rhesus macaques [37], gentoo penguins [12]) all appear in service of food-finding. Despite the representational power of explicit communication, implicit communication proves to be a powerful tool for coordination. In the Serengeti, for example, lions use the posture of fellow pride members to stalk prey covertly [48]. Both Wolves [43] and Dolphins [44]—species that frequently communicate vocally—leverage implicit communication during foraging—adjusting their group formation based on the changes of position and orientation of other pack members [25]. This suggests that implicit communication is of significant importance to multi-agent coordination. We argue that emergent communication should similarly consider a full scope of communicative diversity, rather than over-fitting to natural language. For clarity, we represent these examples as a spectrum of communication in Figure 1.

Implicit communication The role of implicit communication in teamwork is discussed extensively throughout the multi-agent systems [8] [16] [20] [21] [41], and human-robot interaction [6] [11] [50] literature. Following Breazeal et al. [6], we define implicit and explicit communication as follows:

Definition 2.1 (Implicit Communication). Information that is inherent in behavior and is exchanged between agents in a non-deliberate manner.

Definition 2.2 (Explicit Communication). Information that is passed deliberately from one agent (the sender) to another (the receiver) through a communicative action (e.g. vocalization, gesture, etc.).
We now introduce our toroidal pursuit-evasion environment and our curriculum-driven strategy for where \( \vec{v} \) is the agent's velocity. At each time-step, each agent observes the environmental state \( s_t = \{ q_{P_1}, ..., q_{P_N}, q_E \} \) and selects its next heading \( \theta_i \) as an action. The chosen heading is pursued at the maximum allowed speed for each agent (\( |\vec{v}_P| \) for pursuer speed, \( |\vec{v}_E| \) for evader speed). To encourage teamwork, we set \( |\vec{v}_P| < |\vec{v}_E| \). We assume the evader policy to be part of the environment, as defined by the following potential-field cost function:

\[
U(\theta_E) = \sum_i \left( \frac{1}{r_i} \right) \cos(\theta_E - \hat{\theta}_i) \tag{3}
\]

where \( r_i \) and \( \hat{\theta}_i \) are the L2-distance and relative angle between the evader and the \( i \)-th pursuer, respectively. This cost is inspired by theoretical analysis of escape strategies in the pursuit-evasion literature \[45\]. Intuitively, it represents a weighted combination of repulsive force from pairs of pursuers, incentivizing the evader to (i) split the largest bisector of any two pursuers; and (ii) run away from each pursuer individually. We provide a complete derivation of this policy in Appendix A.

### Limitations of planar pursuit-evasion

In general, unbounded planar pursuit-evasion can be described by two cases:

- **Case 1:** \( |\vec{v}_P| > |\vec{v}_E| \). The game is solved by a straight-line chase towards the evader and is not interesting from the perspective of coordination.
Figure 2: a) A planar pursuit-evasion game with periodic boundary conditions interpreted as a toroidal pursuit-evasion environment. b) An intuitive comparison of blind search and EWS.

- Case 2: $|\vec{v}_P| \leq |\vec{v}_E|$. The evader has a significant advantage. Pursuers have at most one opportunity to capture the evader and are usually only successful under strict initialization conditions [45].

Lowe et al. [35] addressed this by penalizing agents for leaving the immediate area defined by the camera with negative reward. The evader defined by Equation (3), however, will run away indefinitely in the $|\vec{v}_P| \leq |\vec{v}_E|$ case. To provoke consistent interaction between agents, we extend the planar environment with periodic boundary conditions. One can think of this as playing the pursuit-evasion game on a torus (see Figure 2a).

We argue that toroidal pursuit-evasion is better suited for studying emergent coordination. First, it does not require strict initialization conditions or special rewards. We can initialize pursuers randomly and allow them to construct ad-hoc formations. Second, pursuit is no longer a one-and-done proposition. This reflects the notion that, in nature, predators often do not give up after a single attempt at a prey—they regroup and pursue it again.

Sparse reward
Pursuers receive a large positive reward ($r(s_t, a_t) = 50.0$) for capturing the prey and a small negative reward ($r(s_t, a_t) = -0.1$) otherwise to incentivize efficient pursuit. Though the negative component of the reward is experienced every time-step, positive reward is experienced sparsely at the end of successful trajectories. Also note that reward is shared—i.e. a pursuer $P_j$ receives a positive reward signal even when another pursuer $P_k$ is the one to capture the evader.

3.2 Curriculum-driven learning

Our goal is to learn multi-agent coordination with decentralized training and sparse reward. Though setting $|\vec{v}_P| \leq |\vec{v}_E|$ is important for studying coordination, it creates a difficult learning problem. In the early stages of training, each pursuer’s action selection is determined by the randomly initialized weights of its policy network. The chance of $N$ slower pursuers capturing the evader defined in Equation (3) with random actions is extremely low. Therefore, so is the likelihood of the pursuers encountering a positive reward signal with which they can improve their policies. This issue is exacerbated by a second challenge: non-stationarity. Non-stationarity is caused by multiple agents performing policy updates simultaneously. In the case of DDPG, non-stationarity 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 leads to higher-variance gradient estimates and unstable learning. Though recent advances in “centralized training, decentralized execution” stabilize training in such cases [35], they violate our goal of decentralized learning.

Instead, we introduce a curriculum-driven method to address the challenges of decentralized learning in sparse reward environments. We assume each pursuer $P_i$ is equipped with a deterministic action policy $\mu_{\phi_{P_i}}$ and can perform DDPG policy updates. Our approach then combines a curriculum over pursuer velocity $|\vec{v}_P|$ and a technique for collecting valuable experience early on in training, which we call Experience Warm-Starting (EWS).

Curriculum over pursuer velocity
Curriculum learning [3] is a popular technique for solving complex learning problems by breaking them down into smaller, easier to accomplish tasks. We can similarly construct a sequence of increasingly difficult pursuit-evasion games by incrementally
Table 1: Summary of coordination and implicit communication abilities of evaluation strategies

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

lowering the maximum speed of the pursuers. Specifically, we initialize the pursuers to be faster than the evader, then anneal their velocity slowly as training progresses. Fast moving pursuers will quickly learn that a straight-line pursuit strategy is effective. As $|\vec{v}_P|$ decays, the pursuers will need to develop increasingly coordinated strategies to capture the evader. More formally, we define a curriculum over velocity bounds. We set an initial velocity $\vec{v}_0$ such that $|\vec{v}_P| > |\vec{v}_E|$ and anneal $|\vec{v}_P|$ to a final value $\vec{v}_{\text{final}}$ over $v_{\text{decay}}$ episodes:

$$|\vec{v}_P| \leftarrow \vec{v}_{\text{final}} + (\vec{v}_0 - \vec{v}_{\text{final}}) \times \max \left( \frac{(v_{\text{decay}} - i)}{v_{\text{decay}}}, 0.0 \right)$$

where $i$ represents the current training episode. After the velocity decay, we allow training to continue for a pre-defined number of episodes. This creates a second “training phase” that allows the pursuers to tune their action policies at $\vec{v}_{\text{final}}$. We describe this process in more detail in Appendix B.

Experience warm-starting Despite moving at a greater speed, the pursuers still struggle to learn how to capture the evader consistently (see ablation study in Appendix B.2). This is because, upon initialization, action selection with random policy weights is equivalent to blind search. Blind search will waste time pursuing irrelevant motion that does not produce valuable experience. We introduce Experience Warm-Starting (EWS), a simple extension of off-policy learning that enables targeted exploration early in the training process. EWS splits training into two exploration phases that use separate behavior policies $\beta_0$ and $\beta_{\mu}$, respectively. The key to EWS is that we can define $\beta_0$ strategically to collect better experience than a random exploration policy. In our experiments, we define $\beta_0$ as:

$$\beta_0 = -\nabla U_{\text{att}}(q_P, q_E)$$

where $U_{\text{att}}$ is the attractive potential-field defined in Equation (1). Note that $\beta_0$ is a greedy policy that runs directly towards the evader. This strategy is obviously sub-optimal when $|\vec{v}_P| \leq |\vec{v}_E|$, but helps pursuers learn to move in the direction of the evader when $|\vec{v}_P| > |\vec{v}_E|$. In our experiments, we use EWS only during the first phase of the velocity curriculum. After EWS, agents follow the standard DDPG behavior policy $\beta_{\mu} = \mu_{\phi}(s_t) + N$, where $N$ is the Ornstein-Uhlenbeck noise process. An intuitive comparison of blind search to EWS-guided exploration is provided in Figure 2b. In the next section, we use this curriculum-driven variant of DDPG—referred to as CD-DDPG—to study emergent coordination and implicit communication in our pursuit-evasion environment.

4 Results

Our evaluation is motivated by 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 this end, we measure the performance of CD-DDPG in our toroidal pursuit-evasion game against a set of analytical and learned strategies of increasing sophistication. We intend for these strategies to 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:

Greedy: Each pursuer follows the greedy control strategy in Equation (1). 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 a variant of CD-DDPG under partial observability. Rather than receiving the complete environment state, each pursuer $P_i$ receives a private observation $o_t =$
{q_P, q_E} consisting of its own location and the location of the evader. Despite not observing each other, CD-DDPG (Partial) pursuers are capable of coordinating through static role assignment. This is equivalent to the pursuers 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 can therefore be thought of as 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 (3):

\[
F(\theta_i, r_i) = \max_{\tilde{\theta}_i, r_i} \left[ \min_{\theta_E} \left[ U(\theta_E) \right] \right] = \max_{\tilde{\theta}_i, r_i} \left[ \min_{\theta_E} \left[ \sum_i \left( \frac{1}{r_i} \cos(\theta_E - \tilde{\theta}_i) \right) \right] \right]
\]

where \(\tilde{\theta}_i\) and \(r_i\) are the polar coordinates of each pursuer relative the evader. Intuitively, this centralized strategy encircles the evader and cuts off potential bisector escape paths while enclosing the circle. It therefore supports both coordination—pursuers work together to uphold a circular formation—and mild 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.

**Capture success** First, we evaluate capture success as a function of the velocity advantage of the evader. The results are shown in Figure 3. As expected, each method has a high capture success rate when \(|\vec{v}_P| > |\vec{v}_E|\). The Greedy strategy drops off at \(|\vec{v}_P| / |\vec{v}_E| = 1.0\), which is also expected—a greedy strategy should only work when \(|\vec{v}_P| > |\vec{v}_E|\). The Pincer and CD-DDPG (Partial) 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 \(|\vec{v}_P|\) decreases further, pursuers following the CD-DDPG strategy significantly outperform the other strategies in terms of capture success. These results show that CD-DDPG has learned to coordinate more effectively than the other strategies. We cannot, however, verify the role of implicit communication through capture success alone. To appropriately study the possibility that implicit communication emerges, we must quantify the extent of each agent’s influence over its teammates.

**Social influence** To study the role of implicit communication in pursuer performance, we compute the Instantaneous Coordination (IC) score for each of our policies. Section 2 discusses how IC can also be interpreted as a measure of implicit communication. Following Jaques et al. [29], we compute IC empirically as a Monte-Carlo approximation over \(N\) multi-agent trajectories. We average influence across all trajectory steps for each pair of agents and take the maximum score between any agent-agent pair. We also evaluate pursuers that act randomly, which provides a baseline for independent action. The results are shown in Figure 4.

We find that IC is relatively low when \(|\vec{v}_P| > |\vec{v}_E|\). This makes sense, as no coordination is needed to capture the evader. The Pincer pursuers, however, maintain a higher IC score as result of their objective incentivizing a circular formation regardless of velocity. The CD-DDPG strategy...
matches the performance of the Pincer strategy, indicating that CD-DDPG has begun learning to coordinate even in these early stages. As $|\vec{v}_P|$ decreases, CD-DDPG noticeably outperforms the other methods. In fact, the steady increase in IC among CD-DDPG pursuers as velocity decreases indicates that CD-DDPG is achieving increasingly complex coordination. This is a promising sign that implicit communication has emerged within the CD-DDPG team—each pursuer is responding to the movements of its teammates. Finally, we note a somewhat surprising minor increase in coordination for the greedy pursuers at low velocities. This is an IC artifact we dub “phantom coordination” and discuss in the Appendix C.2.

High-influence moments We also report on the percentage of high-influence moments that occur between pairs of agents (see Figure 5a). A high-influence moment is defined as a step in which IC is above the mean IC for the entire trajectory. Similar to previous work [29], we find that influence is sparse in most cases—only between 10-15% percent of trajectory steps exhibit high-influence. The exception, notably, occurs for the CD-DDPG pursuers. 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 of $33 - 35\%$. This suggests that, as $|\vec{v}_P|$ decreases, CD-DDPG pursuers form increasingly highly-coordinated formations and are making split-second decisions based on the movements of their teammates. This points more concretely to the emergence of implicit communication between CD-DDPG pursuers, as the findings closely match 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 [25]—which fit with the definition of implicit communication in Section 2.

Relative position during capture One problem with influence-based measures like IC is that we are required to average over entire trajectories of game-play. This assumes that action selection is equally important, regardless of when in a trajectory the action is performed. In episodic tasks like our pursuit-evasion environment, however, the most important time-steps are those that immediately precede capture. For this reason, we look at the distribution of pursuer locations at the time of capture. We collect 100 trajectories that end in capture from each strategy and compute the distribution of pursuer positions relative to the evader, binned in the range $[0, 2\pi]$. In general, we expect coordinated pursuit to exhibit rotational symmetry during capture. Rotational symmetry suggests that the pursuers have learned strategies which lead them to well-defined “capture points” around the evader. Rotational invariance during capture, on the other hand, indicates that the pursuers have not settled into patterns of attack with concrete roles. The results for this study are shown in Figure 5b.

We find that the Greedy and Pincer strategies yield uniform capture distributions. This is unsurprising for Greedy pursuers, whose pursuit paths are not effected by the paths taken by their teammates. For Pincer pursuers, though the Pincer objective incentivizes pursuers to encircle the evader, it does not constrain the pursuers to specific locations on the circle. This leads to a diverse range of capture positions—a reflection of weaker role 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 >40% of the time, which is greater than that of the CD-DDPG pursuit strategy. However, when taken into context with the results from Figure 3 it is clear that this level of rotational symmetry is not necessarily beneficial to pursuit success overall. In fact, it is an example of over-commitment to role assignment. The pursuers adopt very constrained roles—e.g. "P1 always move left", "P2 always move right", etc—which works well at high velocities (\(|\vec{v}_P|/|\vec{v}_E| \geq 1.0\)), but fails at lower velocities. CD-DDPG pursers balance rotational symmetry and rotational invariance during capture. Each pursuer follows a unique angle towards the evader, but does not commit to that angle completely. This means that the CD-DDPG strategy learns structured coordination while simultaneously allowing pursuers to make dynamic adjustments to their position relative the evader to successfully achieve capture. We posit that this dynamic behavior is evidence of implicit communication.

Beyond the quantitative measures described here, we present qualitative findings in Appendix C.

5 Related work

Emergent communication Emergent communication examines the process by which cooperative agents learn communication protocols as a means to completing difficult coordination tasks [31]. Recent work has shown that MARL agents converge upon useful communication protocols in referential games [23] and can even develop language-like properties such as compositionality [10, 46] and Zipf’s Law [9] when exposed to additional structural learning biases. More recently, this class of techniques has expanded to include complex situated environments [15], high-dimensional observations [14], and the negotiation of belief states for reasoning [19]. Though typically focused on explicit forms of communication, further work has shown that agents can learn to communicate implicitly through actions and gestures [38]. This form of action-space communication [1] is most similar to our work. In fact, we consider the study of implicit communication and its role in coordination to be an important first-step in emergent communication research.

Pursuit-evasion Pursuit-evasion is a classic game setting for studying multi-agent coordination [26]. Though often played on a graph [42], work on continuous-space pursuit-evasion has enabled real-world applications such as unmanned aerial vehicles [51] and mobile robots [13]. Further work has shown that optimal control strategies can be derived from value functions [27] or even learned from scratch in MARL setting [22, 35]. Our work extends this work, studying more concretely the role of implicit communication in pursuit. 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) [50] and communication [53] by the pursuers and is subject to theoretical performance bounds under these conditions [45]. Our work uses this setting to study the emergence of animalistic coordination strategies and is similar to Janosov et al. [28] in that regard, but with an emphasis on learned, rather than hand-defined, strategies.

6 Conclusion

This work connects emergent communication to the spectrum of communication that has evolved in nature. Specifically, it highlights 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 slower moving 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. Limitations: Our method assumes full observability in the environment. Though wolves and dolphins (our motivating example) observe their teammates during hunting, a more general system would be operable in partially-observable environments. Future work will explore communication in partial-observability. Societal impact: Our work is applied to pursuit-evasion games. Learning systems that are trained for pursuit-evasion are susceptible to misuse if applied to surveillance systems or autonomous weapons. To avoid misuse, pursuit-evasion environments should only be used in beneficial applications such as search and rescue.
References

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

[2] R Beckers, S Goss, J Deneubourg, et al. Colony size, communication, and ant foraging strategy. 1989.

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

[4] C Berner, G Brockman, B Chan, et al. Dota 2 with large scale deep reinforcement learning. arXiv preprint:1912.06680, 2019.

[5] C Boesch and H Boesch. Hunting behavior of wild chimpanzees in the tai national park. AJPA, 1989.

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

[7] R Bshary, A Hohner, K Ait-el Djoudi, et al. Interspecific communicative and coordinated hunting between groupers and giant moray eels in the red sea. PLoS Biol, 2006.

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

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

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

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

[12] N Choi, J Kim, N Kokubun, et al. Group association and vocal behaviour during foraging trips in gentoo penguins. Scientific Reports, 2017.

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

[14] A Cowen-Rivers and J Naradowsky. Emergent communication with world models. arXiv preprint:2002.09604, 2020.

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

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

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

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

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

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

[21] P Grover and A Sahai. Implicit and explicit communication in decentralized control. In Allerton Conference on Communication, Control, and Computing, pages 278–285. IEEE, 2010.
[22] N Grupen, D Lee, and B Selman. Low-bandwidth communication emerges naturally in multi-agent learning systems. *arXiv preprint arXiv:2011.14890*, 2020.

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

[24] D Hector. Cooperative hunting and its relationship to foraging success and prey size in an avian predator. *Ethology*, 1986.

[25] J Herbert-Read. Understanding how animal groups achieve coordinated movement. *JEB*, 2016.

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

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

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

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

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

[31] A Lazaridou and M Baroni. Emergent multi-agent communication in the deep learning era. *arXiv preprint:2006.02419*, 2020.

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

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

[34] O Lönnstedt, M Ferrari, and D Chivers. Lionfish predators use flared fin displays to initiate cooperative hunting. *Biology letters*, 2014.

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

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

[37] W Mason and J Hollis. Communication between young rhesus monkeys. *Animal Behaviour*, 1962.

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

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

[40] D Neumann and M Orams. Feeding behaviours of short-beaked common dolphins, delphinus delphis, in new zealand. *Aquatic Mammals*, 2003.

[41] E Pagello, A D’Angelo, F Montesello, F Garelli, and C Ferrari. Cooperative behaviors in multi-robot systems through implicit communication. *Robotics and Autonomous Systems*, 29 (1):65–77, 1999.

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

[43] R Peterson and P Ciucci. The wolf as carnivore. *Wolves: Behavior, Ecology, and Conservation*, 2003.
[44] N Quick and V Janik. Bottlenose dolphins exchange signature whistles when meeting at sea. *Biological Sciences*, 2012.

[45] M Ramana and M Kothari. Pursuit-evasion games of high speed evader. *JINT*, 2017.

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

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

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

[49] D Silver, T Hubert, J Schrittwieser, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint:1712.01815*, 2017.

[50] T Vicsek. Closing in on evaders. *Nature*, 2010.

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

[52] K Von Frisch. The dance language and orientation of bees. 1967.

[53] Y Wang, L Dong, and C Sun. Cooperative control for multi-player pursuit-evasion games with reinforcement learning. *Neurocomputing*, 2020.
A Pursuit-evasion strategies (cont’d.)

A.1 Evader escape strategy explained

The goal of the evader strategy in Equation (3) is to run from pursuers along the maximum bisector between two pursuers. Given pursuer positions \( \{q_{P_1},...,q_{P_N}\} \), we compute polar coordinates:

\[
r_i = d(q_E, q_{P_i})
\]
\[
\tilde{\theta}_i = \text{atan2}(y_{P_i}, x_{P_i})
\]

for each pursuer \( P_i \) relative the evader. Next, we define a potential field that will push the evader towards a bisector:

\[
U(\theta_E) = \sum_i \cos(\theta_E - \tilde{\theta}_i)
\]

Using Ptolemy’s difference formula, we can expand the potential field as:

\[
U(\theta_E) = \sum_i \cos(\theta_E) \cos(\tilde{\theta}_i) + \sin(\theta_E) \sin(\tilde{\theta}_i) = A \cos(\theta_E) + B \sin(\theta_E)
\]

when we plug-in the known \( \tilde{\theta}_i \) values. The function \( U(\theta_E) \) is maximized/minimized for values of \( A \) and \( B \) such that:

\[
\nabla U(\theta_E) = -A \sin(\theta_E) + B \cos(\theta_E) = 0
\]

which simplifies to:

\[
\tan(\theta_E) = \frac{B}{A}
\]

The evader follows the direction of the negative gradient \( -\nabla U(\theta_E) \) and pursues it at maximum speed. Modulating the cost function by \( r_i \):

\[
U(\theta_E) = \sum_i \left( \frac{1}{r_i} \right) \cos(\theta_E - \tilde{\theta}_i)
\]

allows the evader to modify its bisector based on the distance to each pursuer. This helps significantly when the evader is stuck in symmetric formations.

A.2 Evader strategy unit tests

We include a set of “unit tests” that shed light on the evader’s decision-making behavior. We assume \( N = 3 \) pursuers are stationed around the evader at relative angles \( \tilde{\theta}_1, \tilde{\theta}_2, \) and \( \tilde{\theta}_3 \). For simplicity, we initialize the pursuers such that \( \forall i, r_i = 1 \) to negate the effects of radius modulation.

- Case 1: \( \tilde{\theta}_1 = 0, \tilde{\theta}_2 = \pi/2, \tilde{\theta}_3 = \pi \). Pursuers are spaced equally around the upper-half of the unit circle. In this case, the cost minimizer occurs for \( \theta_E = -\pi/2 \) (see Figure 6a).

- Case 2: \( \tilde{\theta}_1 = 0, \tilde{\theta}_2 = \pi/2, \tilde{\theta}_3 = -\pi/2 \). Pursuers are spaced equally around the right-half of the unit circle. In this case, the cost minimizer occurs for \( \theta_E = -\pi \) and \( \theta_E = \pi \) (see Figure 6b). Either solution can be selected to move the evader towards the largest opening.

In general, the cosine function imposes structure on the evader’s objective—it will oscillate between \([-1, 1]\) over a period of \( \pi \), taking on a maximum value of \( U(\theta_E) = 1 \) when the difference between the evader’s heading \( \theta_E \) and the relative angle of a pursuer \( \tilde{\theta}_i \) is zero and a minimum \( U(\theta_E) = -1 \) when \( \theta_E - \tilde{\theta}_i = \pi \). Summing over all \( \tilde{\theta}_i \)’s incentivizes the evader to follow the heading that splits the largest bisector of the pursuers, as shown in the examples.

A.3 Pincer pursuit explained

The Pincer strategy described by Equation (6) is inspired by prior work on theoretical pursuit-evasion [45]. Solving Equation (6) requires optimizing over both \( \tilde{\theta}_i \) and \( r_i \). Fortunately, we can exploit the toroidal structure of the environment to construct an optimization routine that solves for \( \tilde{\theta}_i \) and \( r_i \) discretely. In particular, we can unroll the torus \( k \) steps in each direction to generate \((2k + 1)^2\)
Figure 6: Unit tests simulating one time-step of action selection as dictated by the evader’s cost function. (a) Given pursuers at relative angles $\tilde{\theta}_1 = 0, \tilde{\theta}_2 = \pi/2, \tilde{\theta}_3 = \pi$ the evader will select the heading $\theta_E = -\pi/2$, which minimizes its cost. (b) Similarly, the evader will select the heading $\theta_E = -\pi = \pi$ when pursuers are located at relative angles $\tilde{\theta}_1 = 0, \tilde{\theta}_2 = \pi/2, \tilde{\theta}_3 = -\pi/2$.

Figure 7: The torus environment unrolled $k = 1$ times in each direction. The filled in red circles denote the “active” pursuers that are pursuing the evader at the current time-step, whereas the empty circles represent “inactive” pursuers. We consider only a single evader, located in the center tile.

replications of the current environment state. Rather than solving for optimal $\tilde{\theta}_i$ and $r_i$ values directly, we find the set $P$ of pursuers that maximize Equation (6) across all replications of the environment. We constrain the problem by limiting selections of each pursuer $P_i$ to replications of itself only. This dramatically cuts down the number of possible sets $P$ from $\binom{(2k+1)^2N}{1}$ to $\left(\binom{2k+1}{1}\right)^3$, where $N$ is the number of pursuers in the environment. Thus, we solve Equation (6) via a discrete optimization over each of the $\left(\binom{2k+1}{1}\right)^3$ possible pursuer selections. The resulting set $P$ defines the set of “active” pursuers that will pursue the evader directly at the next time-step. Due to the nature of the evader’s objective function—it is attracted to bisectors and repulsed from pursuers—the maximum $P$ tends to favor symmetric triangular formations. Though this method obviously does not scale well with $N$ and $k$, we found that we are able to find a sufficient maximizer with low values of $k$ (i.e. $k = 1$ in our experiments). The replication process is shown for
Figure 8: a) A visualization of the curriculum over pursuer velocity. Training phases are represented by a different color lines. b) Results from the ablation study of curriculum-driven learning early in the training process. Note that pursuer velocity is decreasing over time, which is why learning is characterized not by improved reward, but rather by sustained reward. Alternative training regimes fail to achieve as high of a level of performance in such a short number of episodes.

the $k = 1$ case in Figure 7. Note that we discriminate between “active” pursuers—i.e. those $P_t \in P$ pursuing the evader at the current time-step—from “inactive” pursuers.

### B Curriculum-driven learning

In this section, we highlight the benefits of our proposed curriculum-driven learning strategy, showing through an ablation study that it greatly improves the quality of experience an agent obtains early in the training process. We then discuss self-play as an alternative curriculum learning paradigm. Finally, we present a formal algorithm for CD-DDPG.

#### B.1 Velocity curriculum training phases

The velocity curriculum introduced in Equation (4) defines two separate “training phases”—one in which $|\vec{v}_P|$ decays linearly over $v_{\text{decay}}$ episodes; and one in which $|\vec{v}_P|$ remains constant, allowing the pursuers to tune their action policies at the current velocity level. We start training at a pursuer velocity ratio of $|\vec{v}_P|/|\vec{v}_E| = 1$.2 and anneal $|\vec{v}_P|/|\vec{v}_E|$ by 0.1 over $v_{\text{decay}}$ episodes. At the beginning of each phase after the first, the weights of each agent’s policy and action-value network are copied over from the final checkpoint of the previous training phase. A visualization is shown in Figure 8a.

#### B.2 Ablation study: warm-starting

To isolate the performance improvements of our method, we perform ablations with alternative training strategies. We compare CD-DDPG to two alternatives: No Warm-Start, which follows a curriculum over pursuer velocity (as described in Section 3.2), but does not leverage Experience Warm-Starting; and No Curriculum, which uses neither EWS nor a velocity curriculum during training. We trained each method for 7500 episodes in our pursuit-evasion environment—representing ~1 day of training time—across 10 different random seeds. The warm-up period of training for CD-DDPG was set to 1000 epochs. Results for this ablation study are shown in Figure 8b.

No Curriculum is trained with $|\vec{v}_P|/|\vec{v}_E| = 0.7$ and therefore, unsurprisingly, flat-lines throughout the training process. It is never able to experience positive reward signal, reflecting the difficulty of the pursuit-evasion problem when $|\vec{v}_P| \leq |\vec{v}_E|$. No Warm-Start, trained from an initial velocity ratio of $|\vec{v}_P|/|\vec{v}_E| = 1.2$ downwards, does manage to experience positive reward a handful of times, but not enough to improve the pursuers action policies significantly in such a short period of time. This result reinforces our intuition in Section 3.2—though the pursuers are moving at a greater speed than the evader, the action selection of randomly initialized policy networks still has a difficult time
Figure 9: Qualitative results from the pursuit-evasion experiment. (a) The pursuers coordinate to capture the evader, displaying positional shifts indicative of implicit communication. (b) Top: A diagram of dolphin foraging strategies documented in [40]. Middle: Photos of wolves coordinating while hunting, as shown in [39]. Bottom: The learned behavior of our multi-agent system.

capturing the sophisticated evader. CD-DDPG, by leveraging a strategic behavior policy during warm-up, is able to capture valuable experience 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 EWS and our velocity curriculum is crucial to warm-starting policy learning during the first iteration of the velocity curriculum.

B.3 Why not self-play?

Self-play is a technique for policy training in competitive environments that pits a learning agent against increasingly skilled versions of itself. Self-play defines a self-improvement loop that can be interpreted as an automated curriculum over agent performance—as an agent’s performance improves, so does the performance of its competition. This automated curriculum has enabled some of deep reinforcement learning’s biggest successes, like achieving super-human performance on board games (e.g. Chess, Shogi [49]) and massively open online games (e.g. Dota 2 [4]).

An obvious question for our pursuit-evasion setting is: rather than constructing a curriculum over pursuer velocity, why not define a reward for the evader that is inverse to that of the pursuers, such as:

$$r(s_t, a_t) = \begin{cases} 
-50.0 & \text{if evader captured} \\
0.1 & \text{otherwise} 
\end{cases}$$

and train both the pursuers and evader in a self-play fashion. The primary reason is that, though self-play suffices as a curriculum learning strategy, it creates a dependency between the emergent coordination of the pursuers and the learning process of the evader. This can be problematic if, for example, the evader falls into a local minimum and settles on a sub-optimal escape strategy. This would in turn impact the coordination strategy of the pursuers.

Our approach of defining a potential-field evader strategy creates a more controlled setting for studying pursuer coordination. In our environment, the emergent behavior of the pursuers is solely a product of their own coordination.

B.4 Algorithm

We present CD-DDPG in algorithmic form in Algorithm 1. Though we focus here on DDPG, we note that our strategy is applicable to any off-policy learning algorithm.
We perform post-hoc qualitative analysis of CD-DDPG trajectories. In the trajectories, the pursuers
pursuit at low velocities. In particular, we examine the IC score of the random pursuers.

C.2 Phantom coordination

In Section 4, we describe “phantom coordination” as independent action that is falsely perceived as

C Qualitative results

C.1 Possible emergence of implicit communication

We perform post-hoc qualitative analysis of CD-DDPG trajectories. In the trajectories, the pursuers
appear to adjust their position slightly in response to the movements of fellow pursuers as they close
in on the evader (Figure 8). Moreover, it seems that the pursuers occasionally move away from
the evader—something a less coordinated strategy would not do—to maintain the integrity of the
group formation. This explains the performance difference between CD-DDPG and the competing
analytical strategies, as the potential-field pursuers have no basis for making small-scale adaptive
movements. We interpret these results as encouraging evidence that implicit communication has
emerged amongst CD-DDPG pursuers.

C.2 Phantom coordination

In Section 4 we describe “phantom coordination” as independent action that is falsely perceived as
coordination from the perspective of Instantaneous Coordination. Phantom coordination appears in
the greedy pursuit strategy, where we see the IC score for greedy pursuers increase slightly as the
velocity of the pursuers decreases. This is counter-intuitive because each greedy pursuer ignores the
behavior of its teammates. We would expect the IC score for greedy pursuers to remain flat, mirroring
the IC score of the random pursuers.

To diagnose phantom coordination in our environment, we perform qualitative analysis of greedy
pursuit at low velocities. In particular, we examine N = 3 pursuers as they chase an evader at a
speed of $|\vec{v}_p|/|\vec{v}_E| = 0.4$. The pursuers have no chance of successfully capturing the evader, as
evidenced by their capture success performance at this velocity in Figure 3. However, we find that the
straight-line chase patterns of greedy pursuers form temporary triangular patterns around the evader.
In Figure 10 the greedy pursuers form an ad-hoc triangular formation around the evader for a duration
of 60 time-steps. The leftmost pursuer lies equidistant from the evader around the periodic boundaries
and iterates between moving leftward and rightward, depending on which direction creates a shorter

Algorithm 1 Curriculum-Driven DDPG for a single agent

To diagnose phantom coordination in our environment, we perform qualitative analysis of greedy
pursuers as they chase an evader at a speed of $|\vec{v}_p|/|\vec{v}_E| = 0.4$. The pursuers have no chance of successfully capturing the evader, as
evidenced by their capture success performance at this velocity in Figure 3. However, we find that the
straight-line chase patterns of greedy pursuers form temporary triangular patterns around the evader.
In Figure 10 the greedy pursuers form an ad-hoc triangular formation around the evader for a duration
of 60 time-steps. The leftmost pursuer lies equidistant from the evader around the periodic boundaries
and iterates between moving leftward and rightward, depending on which direction creates a shorter

Algorithm 1 Curriculum-Driven DDPG for a single agent

1. Define initial velocity $\vec{v}_0$, final velocity $\vec{v}_\text{final}$, decay period $v_\text{decay}$, and warm-up episode threshold $W$
2. Initialize actor $\mu_\phi(s)$ with parameters $\phi$ and target actor $\mu_\phi'$ with weights $\phi' \leftarrow \phi$
3. Initialize critic $Q_\omega(s, a)$ with parameters $\omega$ and target critic $Q_\omega'$ with weights $\omega' \leftarrow \omega$
4. Initialize replay buffer $D$
5. Initialize random process $N$ for exploration
6. Receive initial state $s_1$
7. for $t = 1$ to $T$
8.   if $i \leq W$ then
9.      Sample action $a_t = \beta_0(s_t)$
10.     else
11.        Sample action $a_t = \mu_\phi(s_t) + N$
12.   end if
13.  Execute $a_t$, observe reward $r_t$ and new state $s_{t+1}$
14.  Store transition $(s_t, a_t, r_t, s_{t+1})$ in $D$
15.  Sample random minibatch of transitions $\{(s_i, a_i, r_i, s_{i+1})\}_{i=1}^N$ from $D$
16.  Compute TD-target: $y_i = r_i + Q_\omega'(s_{i+1}, \mu_\phi'(s_{i+1}))$
17.  Update critic by minimizing the TD-error: $L(\omega) = \frac{1}{N} \sum_i (y_i - Q_\omega(s_i, a_i))^2$
18.  Update actor using the sampled policy gradient: $\nabla_\phi J(\phi) \approx \frac{1}{N} \sum_i \nabla_\phi \mu(s) \nabla_a Q_\omega(s, a)$
19.  Update target networks: $\phi' \leftarrow \tau \phi + (1 - \tau) \phi'$, $\omega' \leftarrow \tau \omega + (1 - \tau) \omega'$
20. end for
21. Step velocity curriculum: $|\vec{v}_p| \leftarrow \vec{v}_\text{final} + (\vec{v}_0 - \vec{v}_\text{final}) \ast \max((v_\text{decay} - i)/v_\text{decay}, 0.0)$
22. end for
Figure 10: Snapshots from a 60 time-step trajectory in which phantom coordination appears. Though the pursuers are following independent greedy strategies, their actions produce a triangular formation that is perceived as coordination by the IC performance measure.

line to the evader. The other two pursuers approach the evader from above and below, respectively. This behavior causes the evader to move in a zig-zag pattern in the center of the triangle until a large opening appears, through which the evader can escape.

This behavior leads to phantom coordination because IC is computed between two consecutive time-steps and averaged over whole trajectories. This means that, in the case of the greedy pursuers, IC scores for independent actions are averaged together with subsets of each trajectory that consist of seemingly highly coordinated behavior.

D Experiment details

**Potential-field hyperparameters** The Greedy potential-field defined by Equation (1) is subject to a single hyperparameter $k_{\text{att}}$, which defines the strength of the attractive pull of the agent towards its goal. We set $k_{\text{att}} = 1.5$ in all of our experiments.

**Policy learning hyperparameters** Actors $\mu_{\phi}$ are trained with two hidden layers of size 128. Critics $Q_{\omega}$ are trained with three hidden layers of size 128. We use learning rates of $1 \times 10^{-4}$ and $1 \times 10^{-3}$ for the actor and critic, respectively, and a gradient clip of 0.5. Target networks are updated with Polyak averaging with $\tau = 0.001$. We used a buffer $D$ of length 500000 and sample batches of size 512. We used a discount factor $\gamma = 0.99$. All values are the result of standard hyperparameter sweeps.

**Experiments** The models for all experiments were trained following Algorithm 1 for 50000 episodes of 500 steps. We reserve $W = 1000$ episodes for warm-up, during which the EWS behavioral policy $\beta_0$ is followed. Pursuer velocity is decayed over $v_{\text{decay}} = 15000$ episodes. Test-time performance, as in Figures 3, 4, and 5, is averaged across 100 independent trajectories with separate random seeds. For the ablation in Figure 8b, each method was trained for 7500 episodes across 10 random seeds. All experiments leveraged an Nvidia GeForce GTX 1070 GPU with 8GB of memory.

**Assets** The toroidal pursuit-evasion environment used in this work is an extension of the pursuit-evasion environment introduced by Lowe et al. [35]. The original environment is open-sourced on Github under the MIT license. We cited the authors accordingly in the main text. New assets (i.e. the new pursuit-evasion environment) are included in the supplementary material. None of the assets used in this work contain personally identifiable information or offensive content.