MULTI-AGENT REINFORCEMENT LEARNING

Lesson 3: Learning Dynamics – The Evolutionary Game

JULIA WANG

It is not the strongest of the species that survives, nor the most intelligent that survives. It is the one that is the most adaptable to change.

THE DANCE OF AGENTS: A STORY OF CONSTANT MOTION

Single-Agent Learning is a Solo Performance

- The world is a static stage.
- The agent learns steps to a fixed rhythm.
- Goal: Perfect one routine.

Multi-Agent Learning is a Group Dance

- Every dancer's move changes the dance.
- The rhythm changes with every step.
- Goal: Learn to adapt and coordinate.

Today, we learn the choreography of this dance: the dynamics of multi-agent learning.

RECAP & AGENDA

Previously On MARL...

- We explored fundamental architectures:
  - CTCE: The God Controller
  - CTDE: Unified Command
  - DTDE: Total Anarchy
- These are the "blueprints" of our agents.

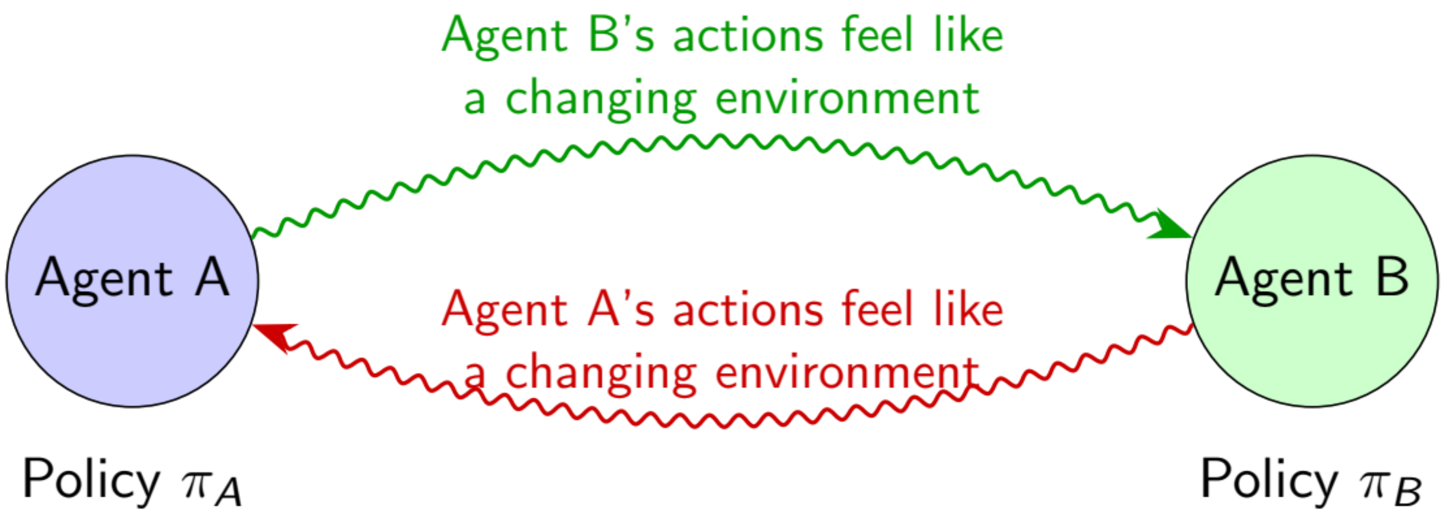
Today's Mission

- Now we see how these agents *evolve*.
  1. **Self-Play**: Creating a perfect sparring partner.
  2. **Policy Gradients**: The math of social influence.
  3. **Mean Field Theory**: Taming the chaos of the crowd.
- **Live Demo!**

THE CORE CHALLENGE: LEARNING ON SHIFTING SANDS

The Central Problem in MARL

From any single agent's perspective, the environment is a **moving target**. As other agents learn and change their policies, the optimal policy for our agent also changes.



The Billion-Dollar Question

How do we achieve stable learning when the "correct" answer is always changing?

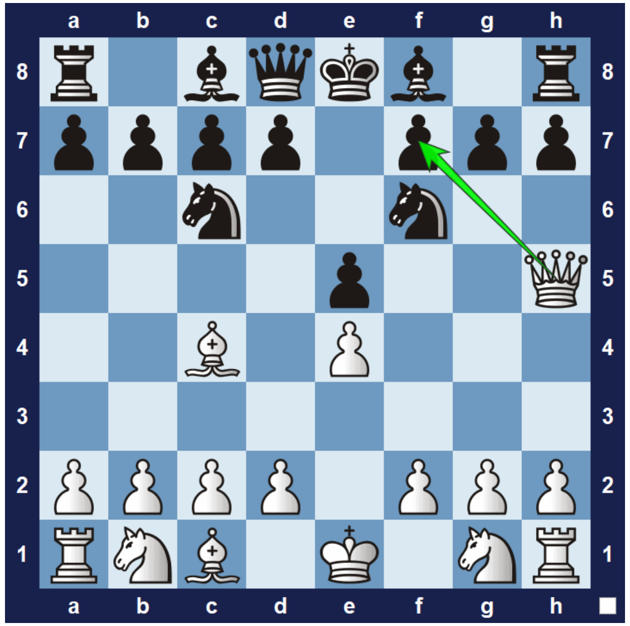
SECTION 1: SELF-PLAY

Fighting Your Own Shadow

THE PROBLEM: OVERFITTING TO A PREDICTABLE FOE

Imagine training a chess AI against a fixed opponent that *only* uses a beginner's 4-move checkmate strategy.

- The AI will quickly learn to counter this one specific attack.
- It becomes the world champion of defending the 4-move checkmate.
- **But it's strategically brittle!** It fails against any other strategy. It has *overfit* to its opponent.



This is the **Red Queen's Fallacy**: you can run faster and faster (get a higher reward), but you're not actually getting smarter if your opponent is standing still.

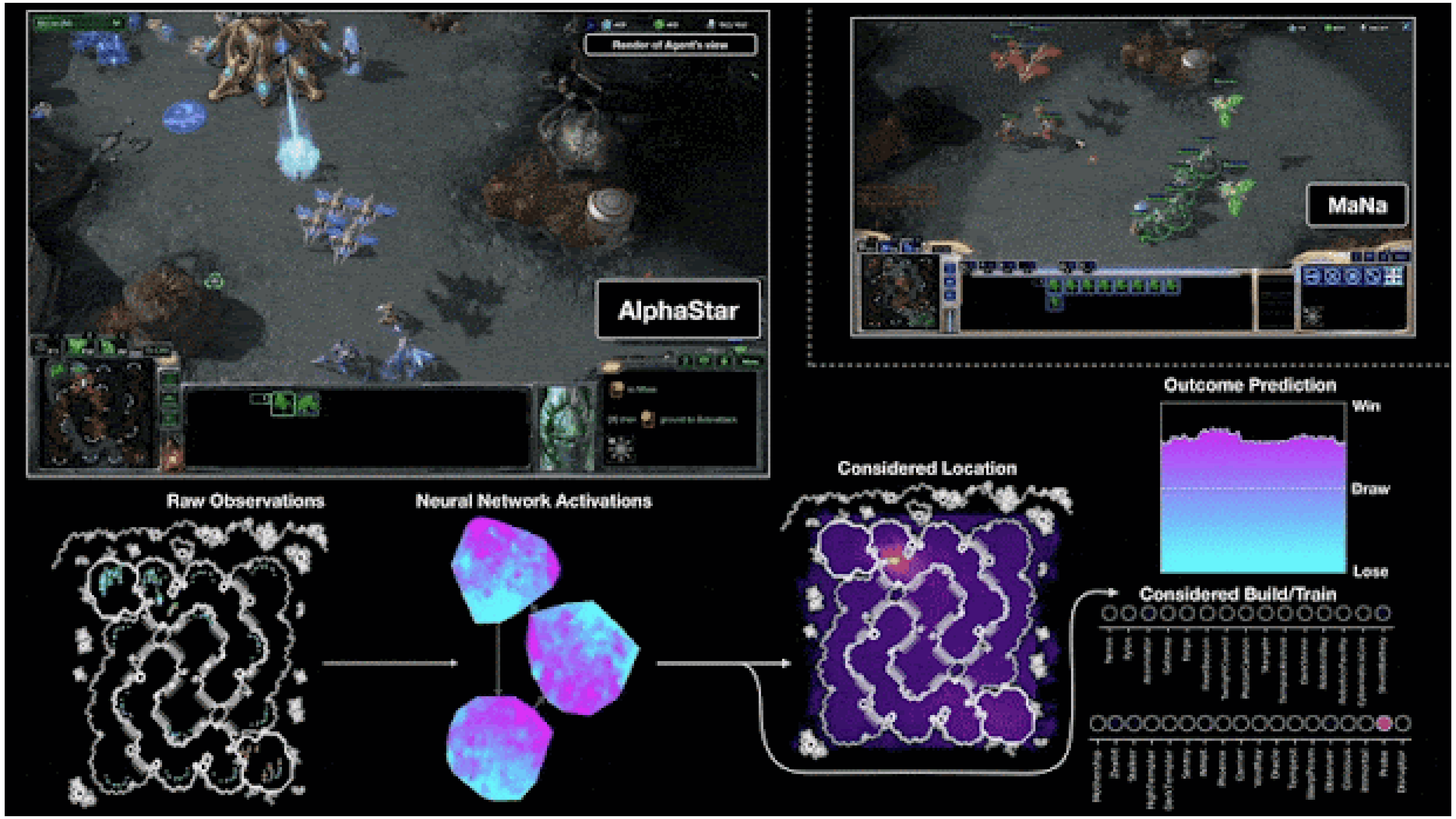
THE SELF-PLAY SOLUTION: A DYNAMIC CURRICULUM

What if your opponent was always a slightly better version of you?

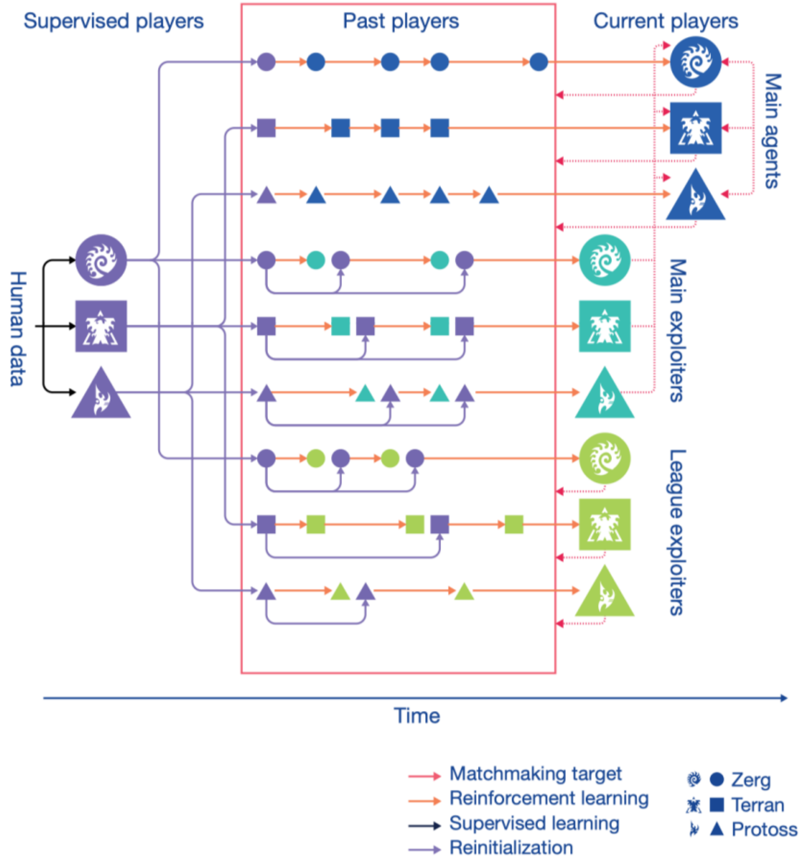
Self-play provides an **autocurriculum**—an automatically generated sequence of increasingly difficult training tasks.

- **Continuous Improvement**: You must constantly adapt to beat your past self.
- **Robustness**: By playing against a diverse pool of your own past strategies, you build a policy that is not easily exploited.
- **Emergent Complexity**: Complex, human-like strategies can emerge from this simple process.

CASE STUDY: ALPHA STAR'S LEAGUE TRAINING



https://deepmind.google/discover/blog/alphastar-mastering-the-real-time-strategy-game-starcraft-ii/



https://xlnwel.github.io/blog/reinforcement%20learning/AlphaStar/

DEEPER DIVE: POLICY-SPACE RESPONSE ORACLES

Self-play isn't just a hack; it's an approximation of a powerful game-theoretic algorithm.

The PSRO Loop:

1. **Start** with a set of initial policies (the meta-strategy or "league").
2. **Compute Best Response**: For each agent, train a new policy that is an optimal "best response" to the current mix of opponent policies in the league.
3. **Add to League**: Add this newly trained best-response policy to the league.
4. **Repeat**: Go back to step 2.

What does this achieve?

This process iteratively builds a strategy set that converges towards a **Nash Equilibrium** of the game. AlphaStar's league is a practical, large-scale implementation of this core idea.

SECTION 2: POLICY GRADIENTS

The Subtle Art of Social Influence

THE CHALLENGE OF INTERDEPENDENT GRADIENTS

In single-agent RL, the policy gradient is straightforward: "If an action led to a good outcome, do it more."

$$\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{\tau \sim \pi_i} \left[\sum_{t=0}^{T} \nabla_{\theta_i} \log \pi_i(a_t|s_t) \quad \underbrace{A_i(s_t, a_t)}_{\text{How good was this action?}} \right]$$

In multi-agent RL, the outcome for agent i depends on everyone's policy (θ_i and θ_{-i}).

The gradient $\nabla_{\theta_i} J_i(\theta_i, \theta_{-i})$ is contaminated by the choices of others! How can we assign credit or blame correctly?

APPROACH 1: INDEPENDENT LEARNING (THE OPTIMIST)

Analogy: Ignorant Dancers

Each dancer tries to perfect their moves in isolation, hoping everyone else does the same.

Method: (IQL, IPPO)

- Each agent treats all other agents as part of the static environment.
- It calculates its gradient $\nabla_{\theta_i} J(\theta_i)$ completely ignoring the fact that θ_{-i} are also changing.
- **Pro**: Incredibly simple and scalable. It's just single-agent RL replicated N times.
- **Con**: Severely violates the stationarity assumption. Often fails to converge, leading to chaotic, unstable policies.

APPROACH 2: CENTRALIZED CRITIC (THE COORDINATOR)

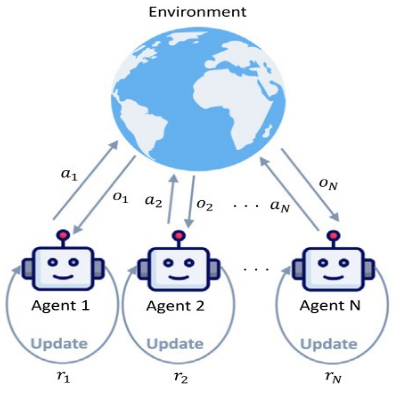
Analogy: The Dance Instructor

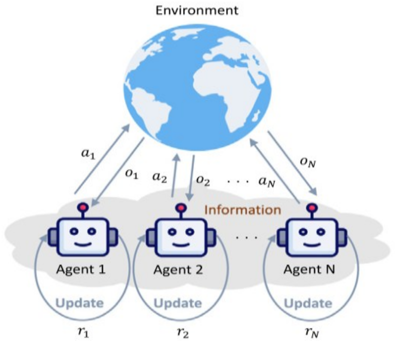
A central instructor watches everyone and gives personalized feedback, but the dancers must perform on their own during the final show.

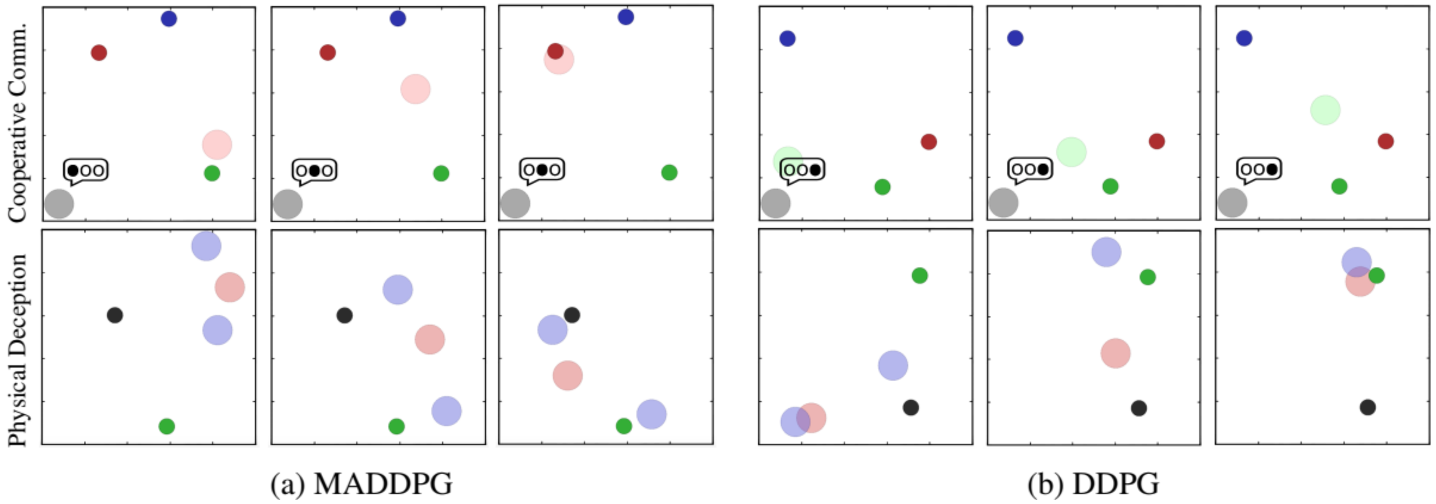
Method: (MADDPG, COMA, MAPPO)

- Follows the Centralized Training, Decentralized Execution (CTDE) paradigm.
- **During training**: A centralized critic sees everyone's observations and actions $(s, a_1, ..., a_N)$.
- This allows the critic to learn an accurate value function $Q_i(s, a_1, ..., a_N)$ and provide a stable, informed gradient to each agent.
- **During execution**: The critic is thrown away. Each agent acts using only its local policy.

CASE STUDY: INDEPENDENT & CENTRALIZED







https://www.sciencedirect.com/science/article/pii/S2949855424000042

https://arxiv.org/pdf/1706.02275

SECTION 3: MEAN FIELD THEORY

From Individuals to Population Trends

THE CURSE OF MANY AGENTS

Problem: The joint action space grows *exponentially* with the number of agents.

- 2 agents, 4 actions each: $4^2 = 16$ joint actions. (Easy)
- 5 agents, 4 actions each: $4^5 = 1,024$ joint actions. (Manageable)
- 10 agents, 4 actions each: $4^{10} > 1,000,000$ joint actions. (Intractable)
- 100 agents... 😵

We need a way to simplify!

Can we approximate the effect of the crowd without modeling every single individual?

MEAN FIELD THEORY: THE ”STATISTICAL AVERAGE” APPROACH

Instead of tracking every agent, track the behavior of the *average* agent.

The core assumption of MFT is that the influence of any single agent on another becomes negligible as $N \to \infty$. What matters is the **collective, average effect** of the population.

- An N-player game is simplified into N parallel 2-player games.
- Each game is played between an agent i and the "mean field" (the average policy of all other agents, $\bar{\pi}$).

HOW MEAN FIELD RL WORKS: THE MATH

The standard Q-function depends on all individual agent actions and states:

$$Q_i(s_i, a_i, s_{-i}, a_{-i})$$

This is approximated in Mean Field Q-learning by taking the expectation over the *average action* $\bar{a}$ from the mean policy $\bar{\pi}$:

$$Q_i(s_i, a_i) \approx \mathbb{E}_{\bar{a}_{-i} \sim \bar{\pi}}[Q_i(s_i, a_i, \bar{a}_{-i})]$$

The Payoff

The Q-function for agent i now only depends on its own state-action and the *mean policy* of its neighbors, drastically reducing complexity.

APPLICATION: SWARM ROBOTICS

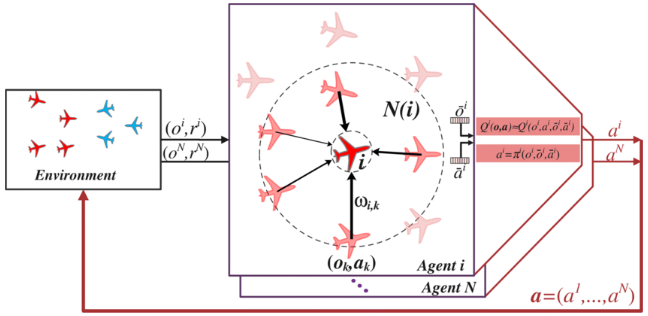
Challenge

- Coordinate thousands of simple drones or robots.
- It's intractable to model every drone-to-drone interaction.

Mean Field Solution

- Each drone doesn't need to know what every other specific drone is doing.
- It only needs to react to the average movement, density, and direction of the swarm in its vicinity.

https://www.azorobotics.com/Article.aspx?ArticleID=4



https://link.springer.com/article/10.1007/s10489-022-03840-6

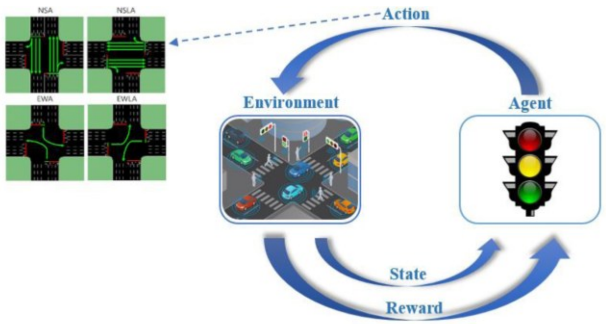
APPLICATION: ECONOMICS & TRAFFIC

Challenge

- Simulate city-wide traffic flow or financial markets with millions of participants.
- The behavior of any one driver or trader is statistically insignificant.

Mean Field Solution

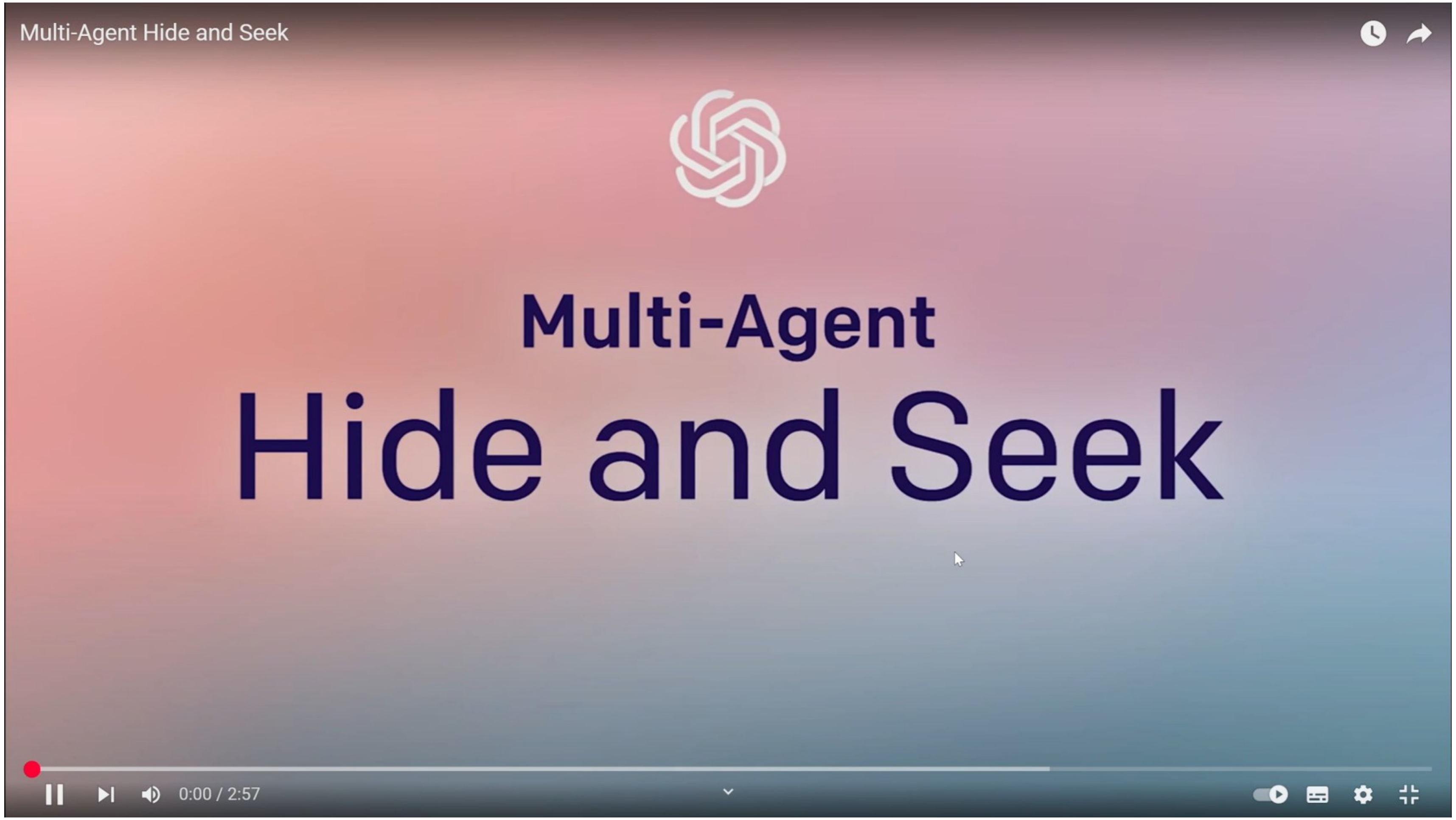
- Model how an individual driver reacts to average congestion levels.
- Model how a trader reacts to average market sentiment (e.g., bull vs. bear market).



https://www.sciencedirect.com/science/article/pii/S1084804522001394

https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2021.749878/full

LIVE DEMONSTRATION



https://www.youtube.com/watch?v=kopoLzvh5jY

Let's watch these learning dynamics in action...

SUMMARY KEY TAKEAWAYS

1. **Learning dynamics are the core of MARL**: Non-stationarity is not a bug, it's a feature of this "evolutionary game."
2. **Self-Play creates a robust learning curriculum**: It forces agents to become robust and general by creating a never-ending arms race (AlphaStar).
3. **Policy gradients require careful coordination**:
   - *Independent Gradients* (The Optimist): Simple, but often unstable.
   - *Centralized Critics* (The Coordinator): Stable and effective, the cornerstone of modern CTDE methods.
4. **Mean Field Theory is the key to massive scale**: It tames the curse of dimensionality by replacing individual interactions with a population average.

HOMEWORK: INDEPENDENT VS. CENTRALIZED PPO

Theoretical Questions

1. Read Lowe, R. et al. (2017). *Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments.*
2. In 3-4 sentences, explain how MADDPG's centralized critic provides a stable learning signal, and why this is not possible for an independent learner.

Practical Challenge

- **Goal**: Complete a MAPPO (Multi-Agent PPO) implementation and compare its performance against a provided IPPO baseline.
- **Environment**: PettingZoo's 'simple_spread', where agents must learn to cover target landmarks.
- **Task**: A Python script with the IPPO baseline and a MAPPO skeleton is provided. Your job is to fill in the missing sections in the `MAPPOAgent`'s `update` method.

NEXT TIME ON MARL…

Lesson 4: Real-World Complexities – From Theory to Practice

We leave the ideal world behind and tackle the messy, practical challenges of real-world MARL.

- **Challenge 1: The Fog of War (Partial Observability)**
  - How do agents make decisions with incomplete information?
  - Using Memory and Attention to see through the mist.
- **Challenge 2: The Art and Science of Communication**
  - From learned "secret handshakes" to efficient, compressed messages.
  - Hierarchical Coordination: The "Manager-Worker" paradigm.
- **Challenge 3: Robustness and Safety**
  - How to build agents resilient to adversarial attacks and noise.
  - Balancing performance with critical safety constraints.

Questions?