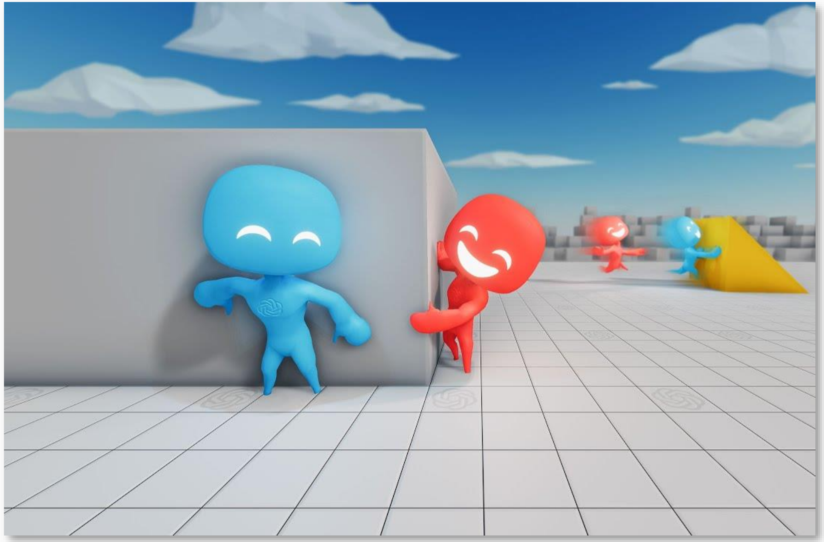
EMERGENT TOOL USE FROM MULTI-AGENT AUTOCURRICULA



<u>Presentation Team:</u> Team 6

3076 Sotiris Ftiakas

3108 Grigoris Barbas

3184 Konstantinos Loizou

3187 Loukas Chatzivasili

3195 Nikolas Petrou

INTRODUCTION

GENERAL ASPECTS

- Simple game of hide-and-seek
- Multi-agent autocurricula
- Reinforcement Learning
- Creators – Researchers:
    - 6 Members of OpenAI
    - 1 Member of Google Brain

Google Brain

OpenAI

HIDE AND SEEK GAME

WHAT DOES THE GAME CONSISTS OF?

2 Teams

Hiders and Seekers

Moveable Objects

Ramps, Boxes and interior walls

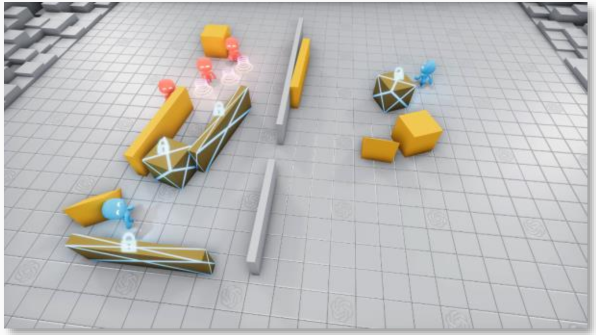
Immovable Objects

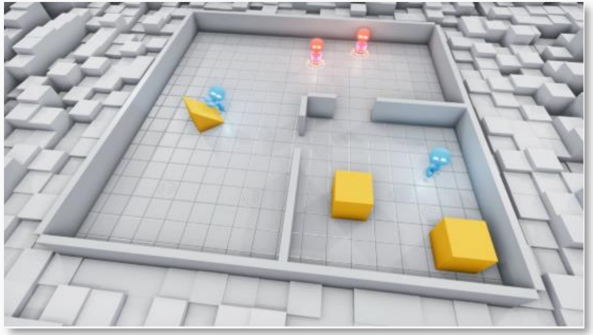
Randomly generated walls and rooms

Preparation phase

Hiders can move
Seekers are immobilized

ENVIRONMENT'S SIMULATION





Physics-based environment (MUJOCO engine)

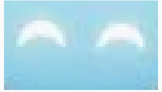
For each episode:

- Hiders: 1-3
- Seekers: 1-3
- Movable objects: 5-11
- Randomly generated static walls and rooms
- No supervision for objects
- Preparation phase: 40% of the episode

AGENTS' SIMULATION

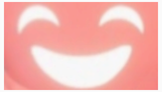
- Simulated as spherical objects
- 135-degree front vision
- Sensors around agents
- Observe other agents
- Team-based visibility reward function

Hiders



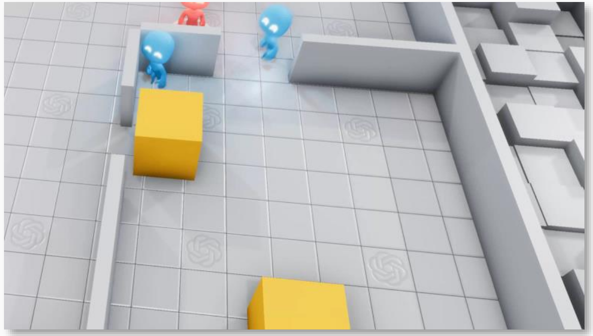
+1 if all hiders are hidden
-1 otherwise

Seekers

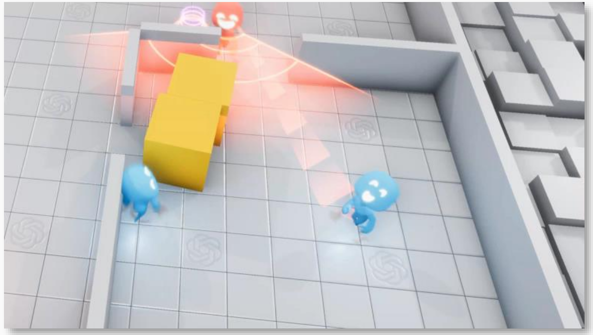


-1 if all hiders are hidden
+1 otherwise

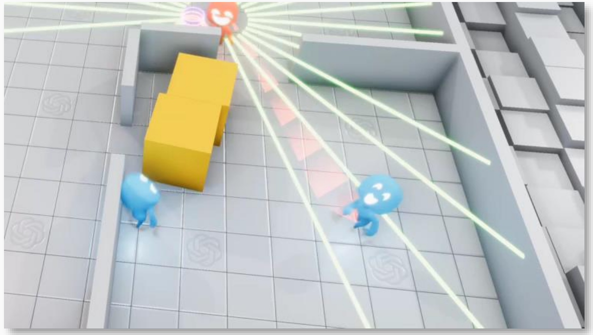
AGENTS' ACTION TYPES



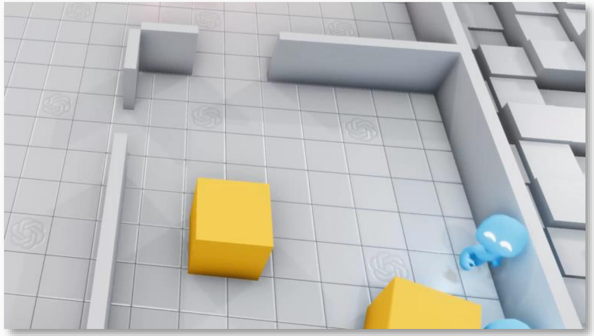
Move in x, y axis and rotate along z-axis



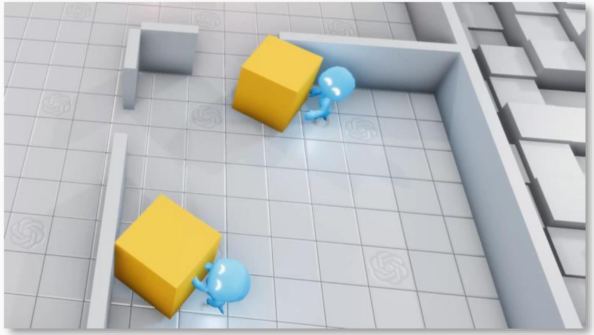
See objects in their line of sight



Sense distance to objects, walls, and other agents using a sensor



Grab and move objects



Lock objects in place

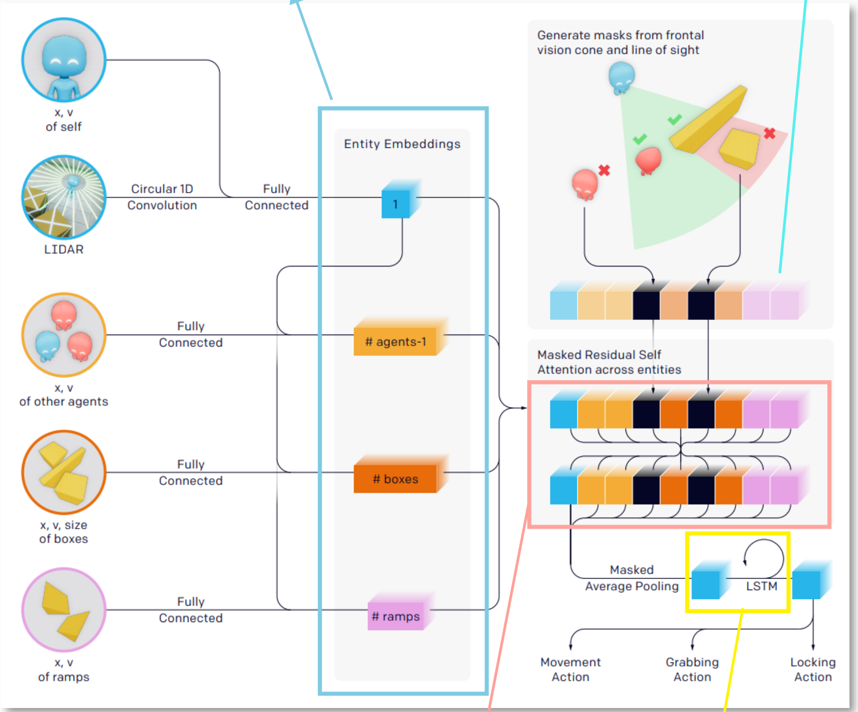
OPTIMIZATION DETAILS

POLICY ARCHITECTURE

- Ego-centric policy
- x, v for state and velocity
- Entities embedded with fully connected dense layers
- Shared weights on same entity types
- Embedded entities passed through residual self-attention block
- Average pooling, action decision from LSTM
- Normalization layer to every hidden layer

Size of embedding layer 128

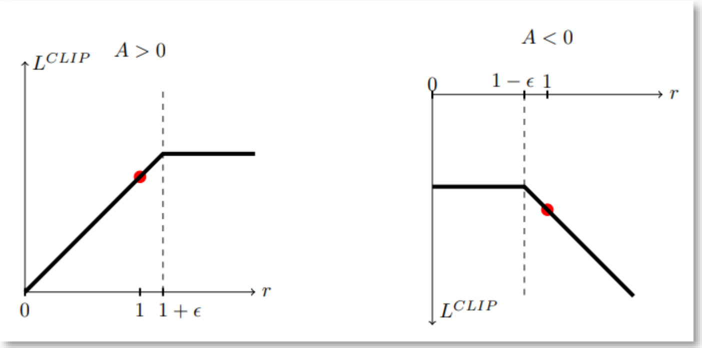
Residual attention layer 4 attention heads of size 32



Size of MLP layer 256

Size of LSTM layer 256

POLICY OPTIMIZATION



- Stochastic Policy
- Proximal Policy Optimization (PPO)
  - Policy Gradient Variant
  - Penalizes large policy changes – prevents instabilities
  - Optimizes the objective:

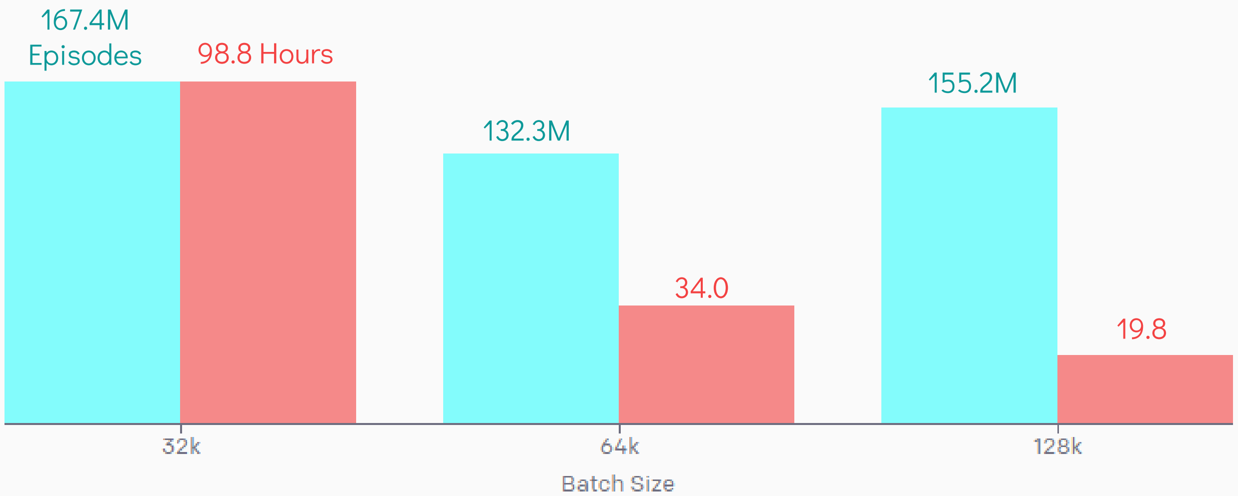
$$L = \mathbb{E}\left[min\left(l_t(\theta)\hat{A}_t, clip\left(l_t(\theta), 1 - \epsilon, 1 + \epsilon\right)\hat{A}_t\right],\qquad l_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{old}\,(a_t|s_t)}$$

- Generalized Advantage Estimation (GAE)
  - Advantage Function

$$\hat{A}_t^H = \sum_{l=0}^{H} (\gamma\lambda)^l \delta_{t+l}, \delta_{t+l} := r_{t+l} + \gamma V\left(s_{t+l+1}\right) - V\left(s_{t+l}\right)$$

LARGE SCALE TRAINING

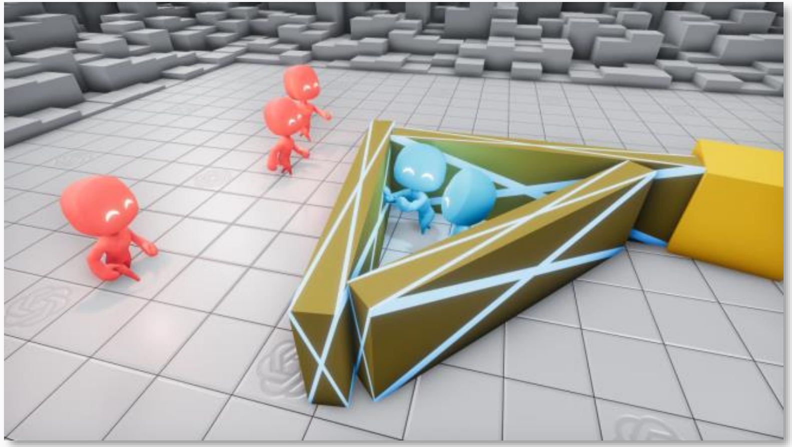
- Critical role in enabling progression
- Larger batch sizes - speeds up convergence
- Default batch size 64,000
- Batch sizes of 16,000 - 8,000 never converged



EMERGENT BEHAVIOR

SIX DISTINCT STRATEGIES

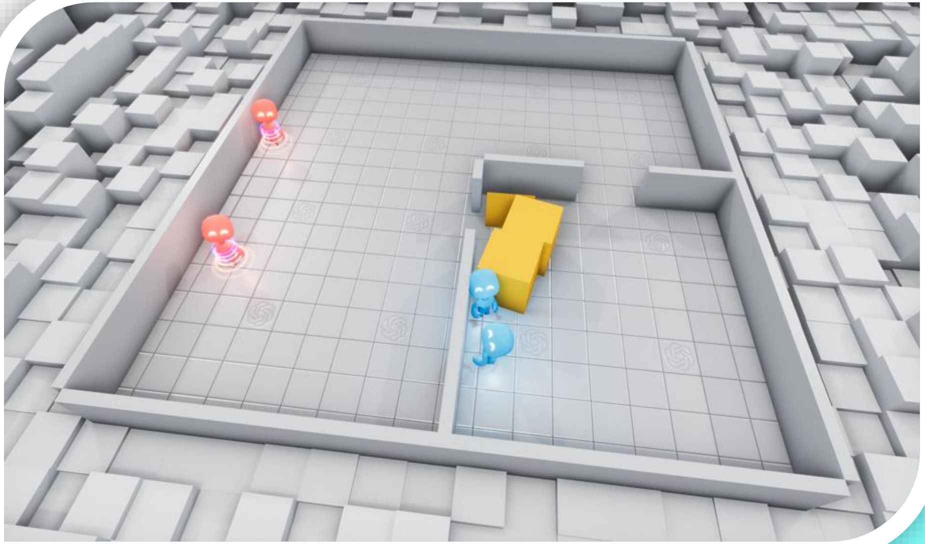
- Run away and chase
- Shelter building
- Ramp use
- Ramp defense
- Box surfing
- Surf defense



- Strategies used by hiders
- Strategies used by seekers

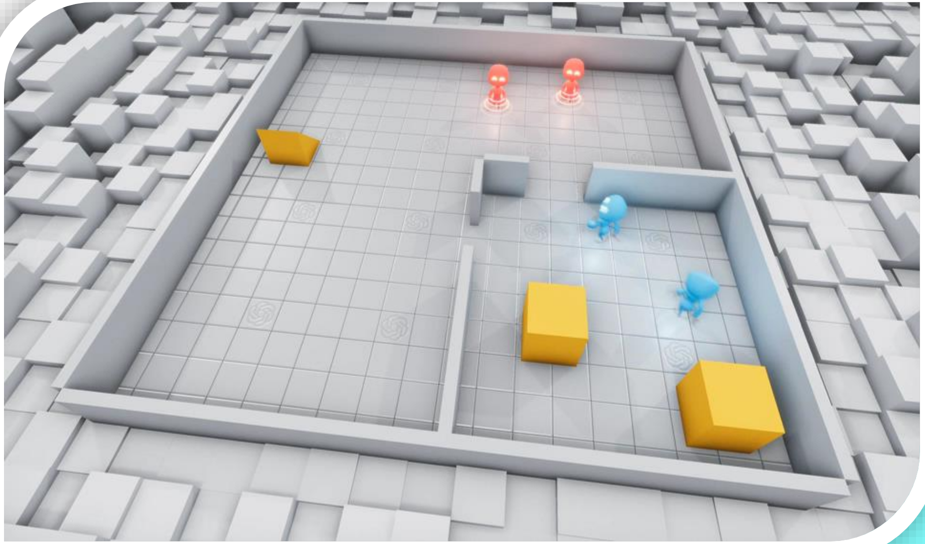
RUN AWAY AND CHASE

- Hiders learn to avoid sight of seekers
- Seekers learn to keep sight of hiders
- Objects are not used



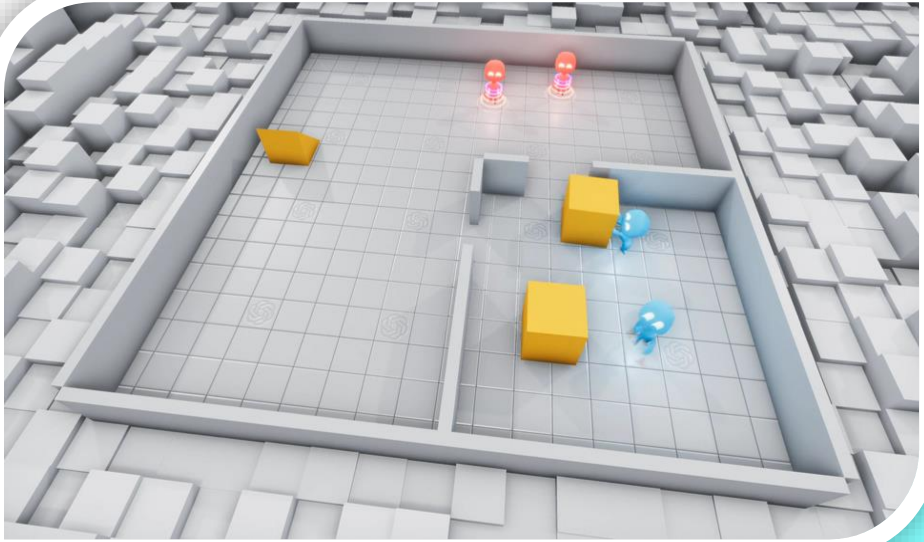
SHELTER / FORT BUILDING

- Learn to modify the environment
- Move boxes to create shelters
- Put boxes against the walls and doors
- Lock boxes in place



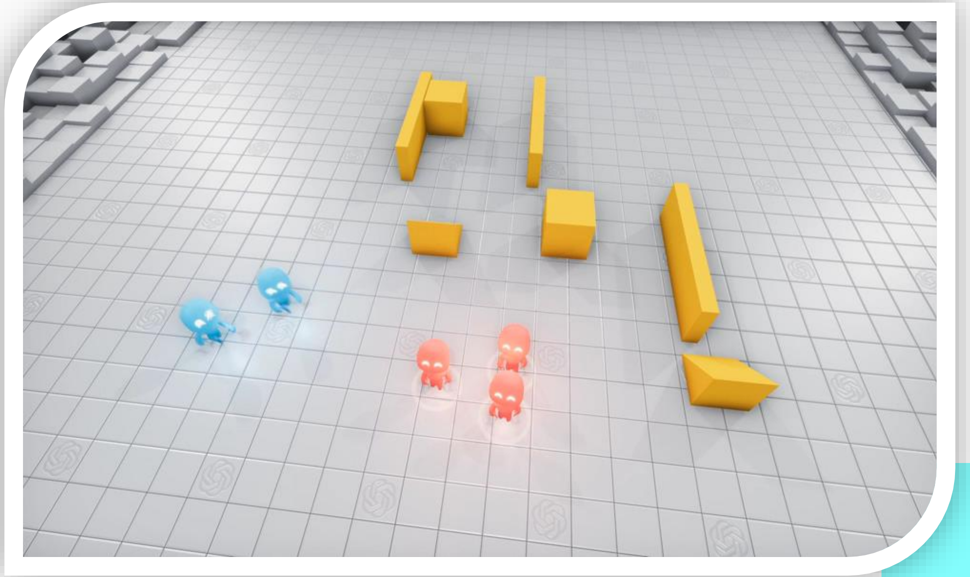
RAMP USE

- Learn to modify the environment
- Move ramps next to shelter
- Climb over ramps and jump in shelter



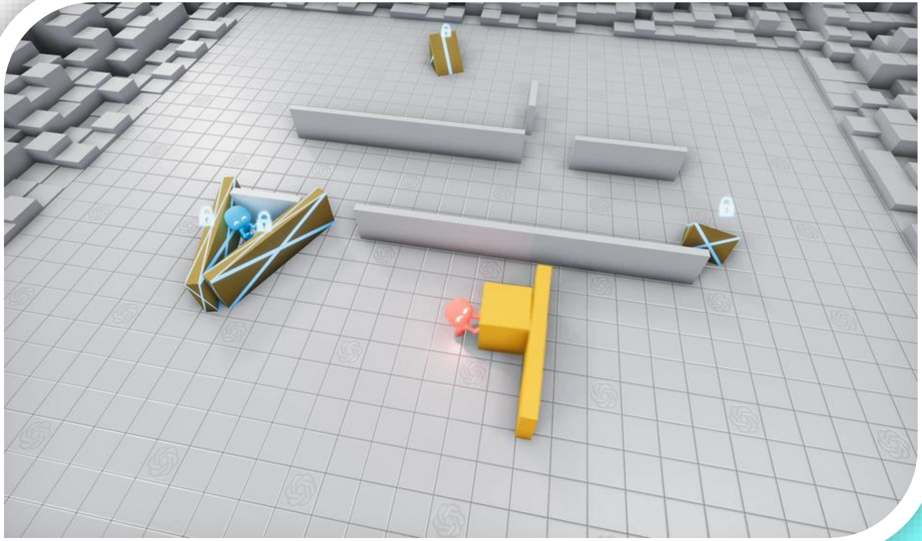
RAMP DEFENSE

- Lock ramps away from the shelter
- Lock ramps inside the shelter



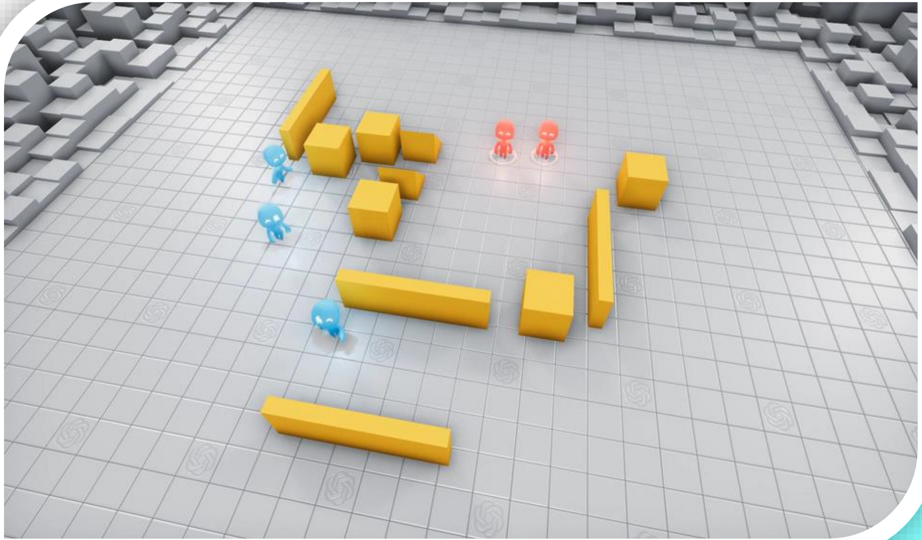
BOX SURFING

- Push boxes next to locked ramps
- Use the ramps to climb on boxes
- Surf boxes to shelter and jump in



SURF DEFENSE

- Lock ramps
- Lock extra boxes to defend surfing
- Create shelter with the remaining boxes



COLLABORATION

- Agents divide the labor
- Hiders pass ramps to one another
- Hiders carry their own box for construction

RANDOMIZATION OVER COMPONENTS

- Reduced randomization results:
  - Fewer stages of evolution
  - Less sophisticated strategies

team size	# boxes	box shape	initial location	walls	emergence
1-3	3-9	random	random	random	6 stages
1	3-9	random	random	random	6 stages
1	7	fixed	random	random	6 stages
1-3	3-9	random	random	no walls	4 stages
1	3-9	random	random	no walls	2 stages*
1	7	fixed	random	no walls	2 stages*
1	7	fixed	fixed	no walls	2 stages

* Hiders run away and use boxes as movable shields.

EVALUATION

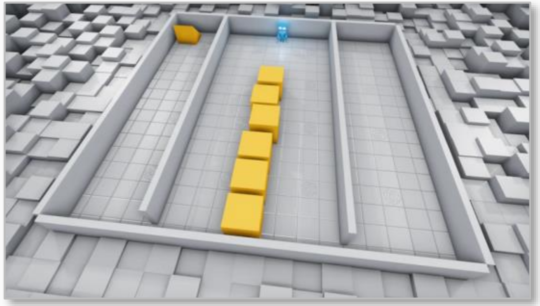
ASSESSMENT

- Assess agent capabilities
- Testing on different domain-specific tasks
- 3 types of agents:
    1. From scratch
    2. Pretrained with Multi-Agent Hide-And-Seek policy
    3. Pretrained with Count-Based Intrinsic Motivation policy
- 5 total benchmark intelligence tests

INTELLIGENCE TESTS

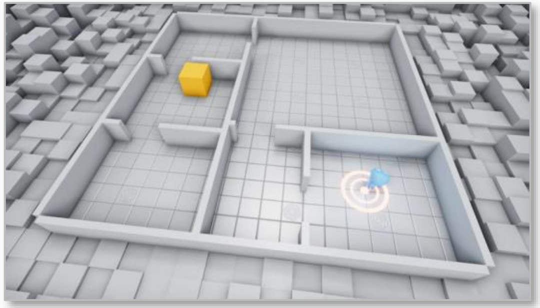
- Cognition and Memory Tests

Object Counting



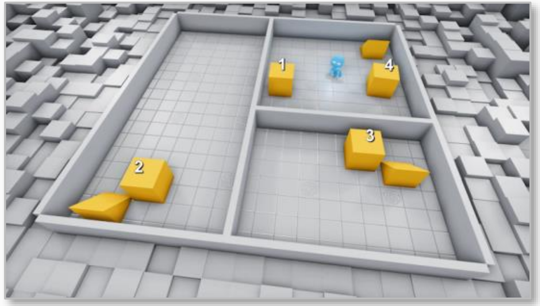
Goal: Memory and sense of object permanence

Lock and Return



Goal: Long-term memory

Sequential Lock

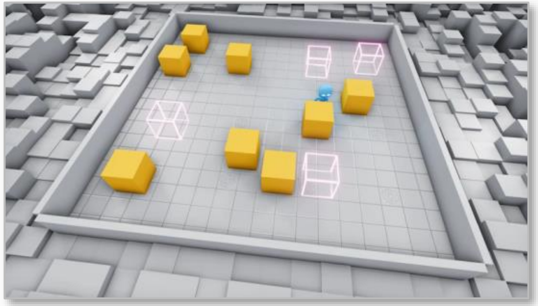


Goal: Lock boxes in a particular order

INTELLIGENCE TESTS

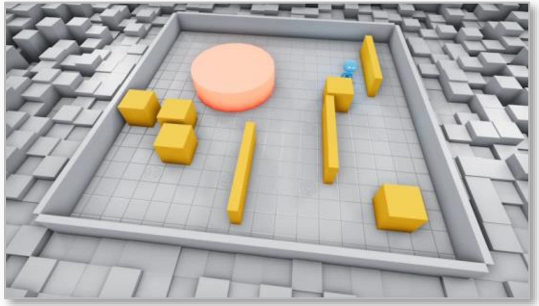
- Manipulation Tests

Blueprint Construction



Goal: Move boxes to the target location

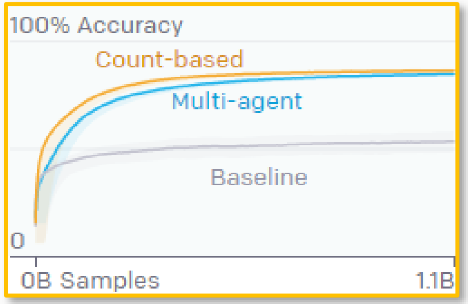
Shelter Construction

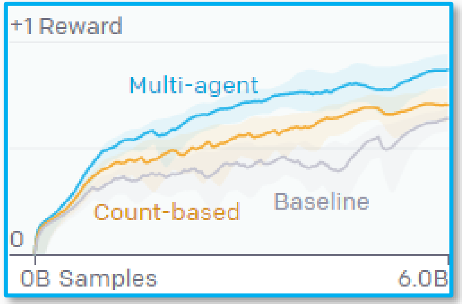


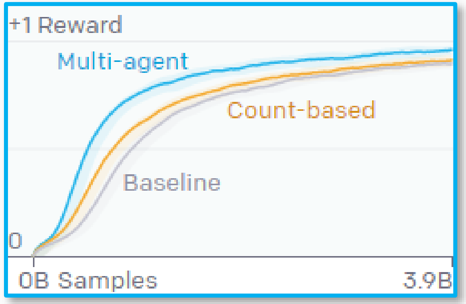
Goal: Construct a shelter around the cylinder

RESULTS

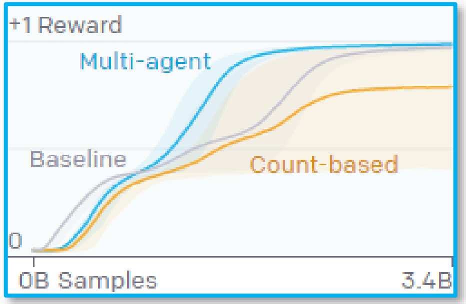
Cognition and Memory Tests

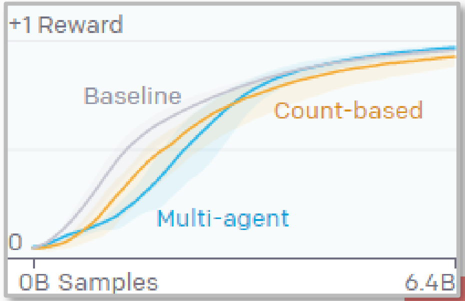






Manipulation Tests

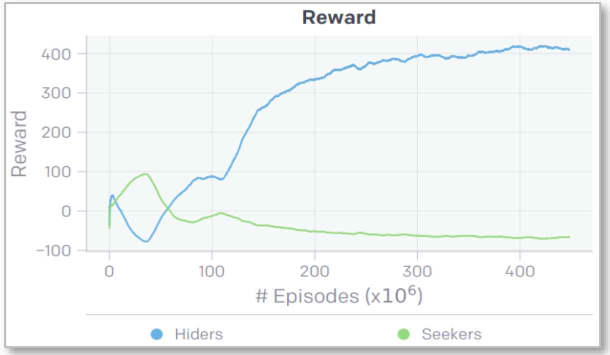


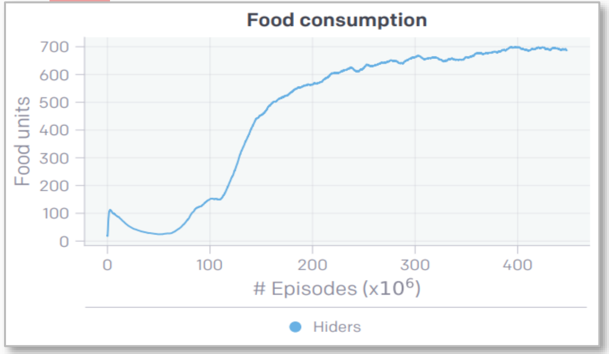


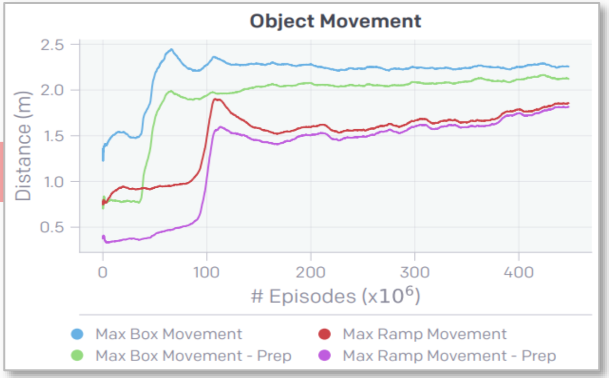
ALTERNATIVE GAME MODE

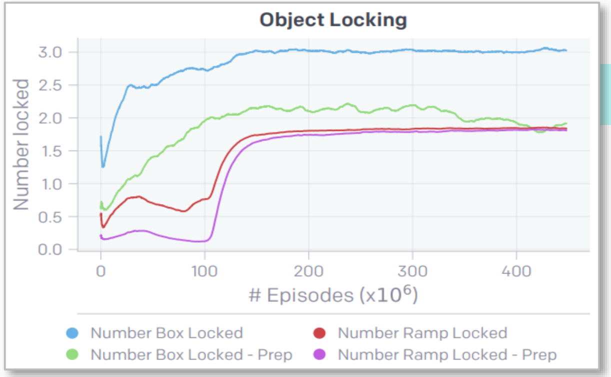
- Secondary objective for Hiders – Food rewards
- Conditions:
  - ❑ Eat food after preparation phase only
  - ❑ Food should be close and visible
  - ❑ No food reward while being seen by Seekers
- Emergent Strategy: Build fort around food location
- 4 levels of skill progression

STATISTICAL GRAPHS





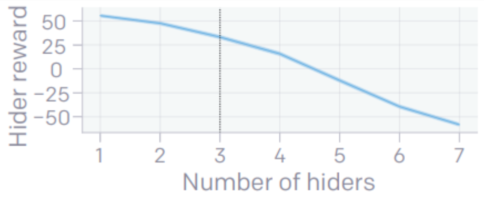


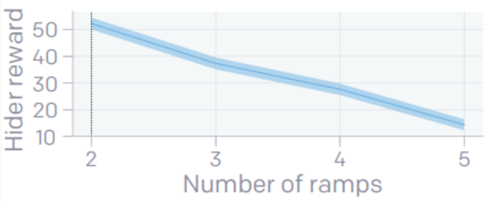


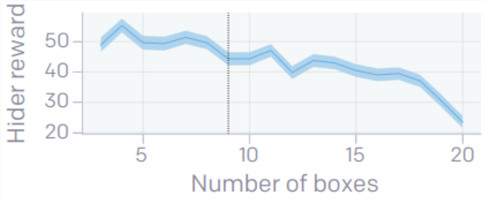
ZERO-SHOT GENERALIZATION

ZERO-SHOT GENERALIZATION

- Trained policies zero-shot generalize to larger environments
- Hider reward as a measure for generalization performance
- Increased hiders – gradual decline in hider reward
- Increased ramps – gradual decline in hider reward
- Increased boxes – stable-slow decline in hider reward







REVIEW

REVIEW

Self-play can lead to emergent autocurricula in agent strategy

Multi-agent autocurricula develop human-relevant skills

Open-sourced environment

SOURCES

- Bowen, B., Ingmar, K., Todor, M., Yi, W., Glenn, P., Bob, M., & Igor, M., *EMERGENT TOOL USE FROM MULTI-AGENT AUTOCURRICULA.* ICLR. (2020)
- John, S., Filip, W., Prafulla, D., Alec, R., Oleg, K., *Proximal Policy Optimization Algorithms,* (2017)
- OpenAI - Emergent Tool Use

THANKS!

Do you have any questions?