Iterated Reasoning with Mutual Information in Cooperative and Byzantine Decentralized Teaming

Sachin Konan*, Esmaeil Seraj*, and Matthew Gombolay

ICLR 2022

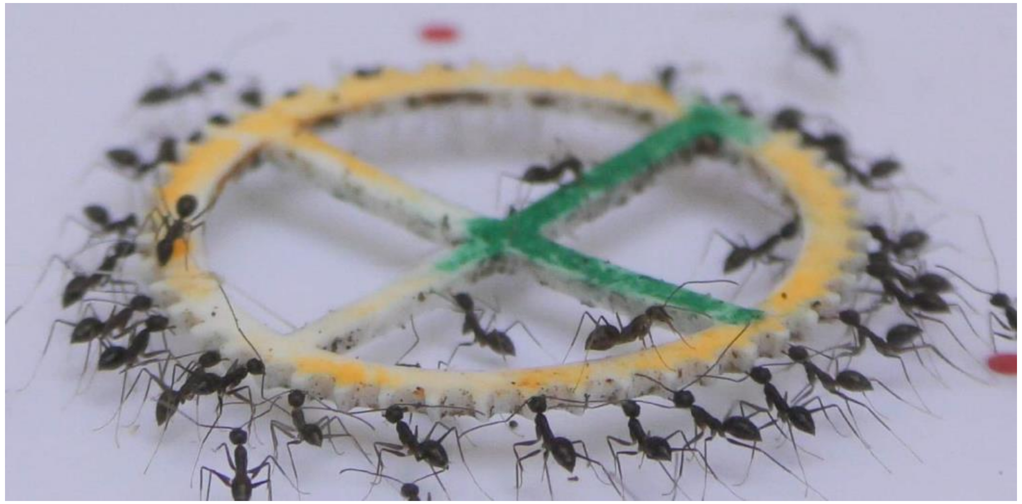
*** Co-first authors (Equal Contributions)**

02/10/2022

Background: Information sharing for multi-agent teaming

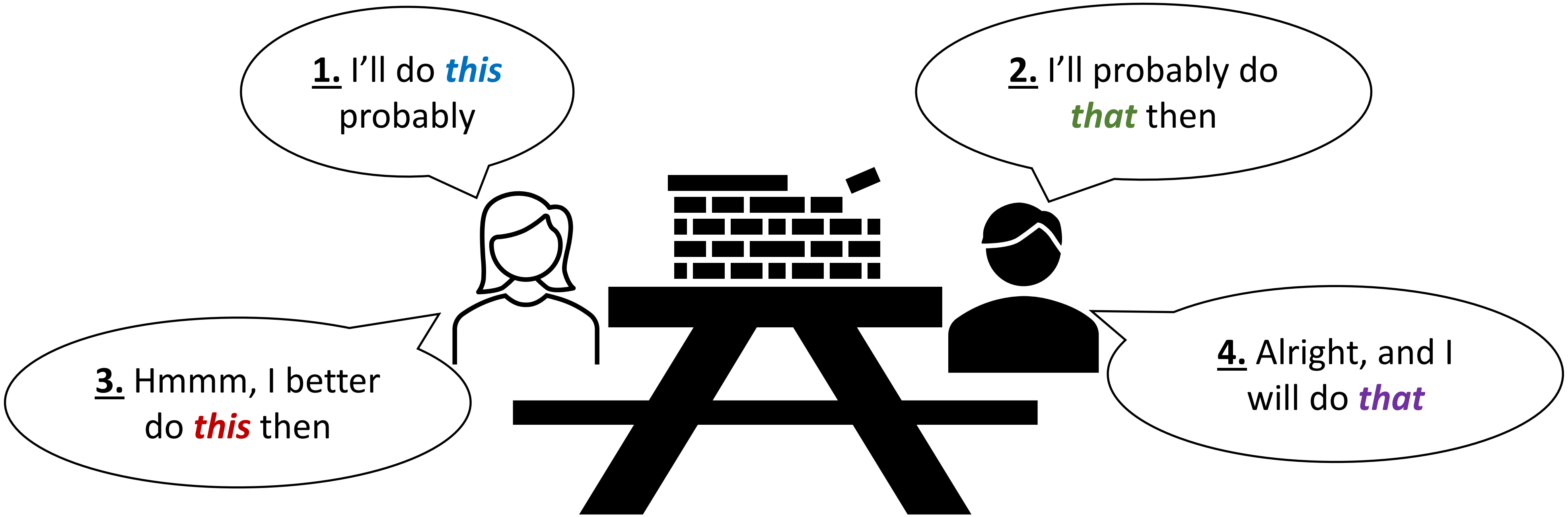
Information sharing and communication, a key feature in building team cognition.

Communication ⇨ Coordination ⇨ Collaboration



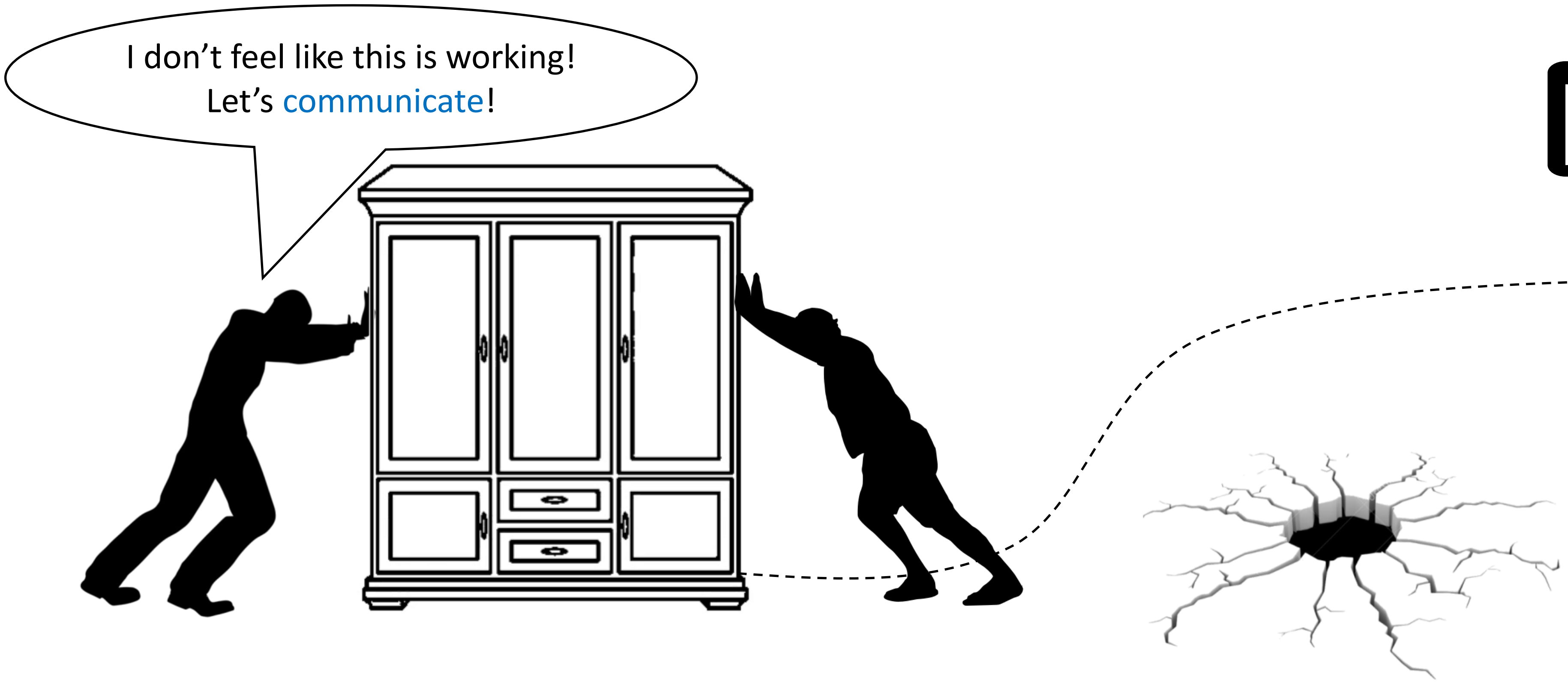
Background: Iterated communication and rationalizability

- High-performing human teams act strategically



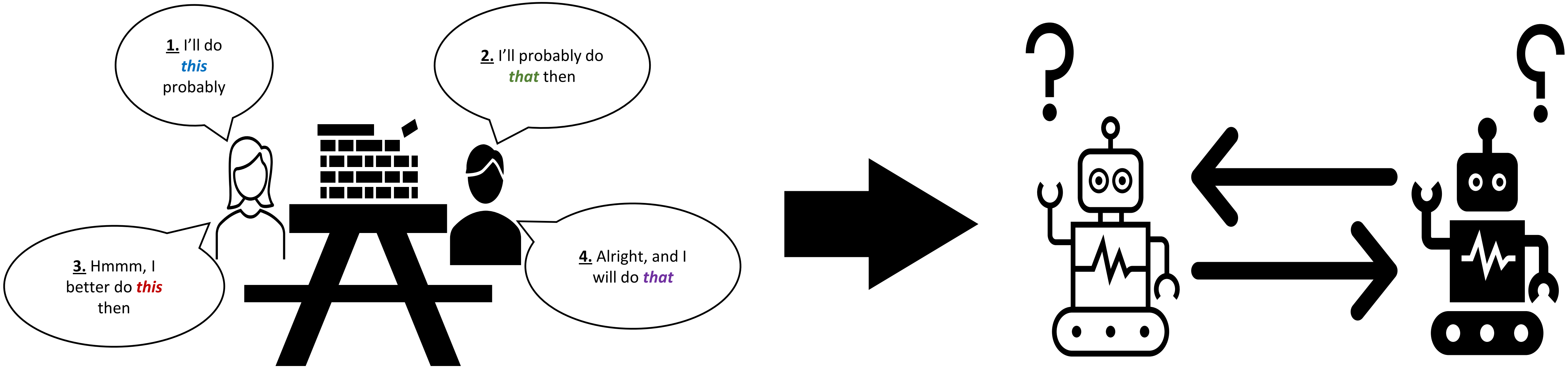
Background: Prior work in multi-agent RL

- Too strong to assume all teammates are perfectly rational in their decision-making!



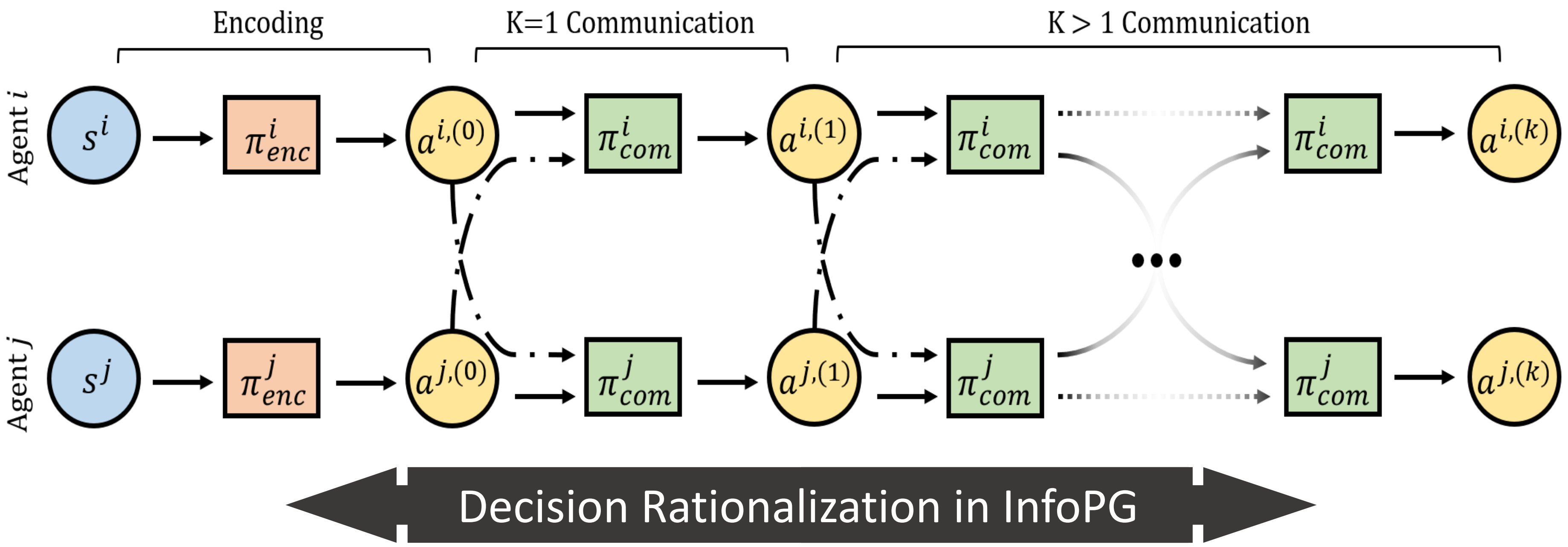
In this paper: Informational Policy Gradient (InfoPG)

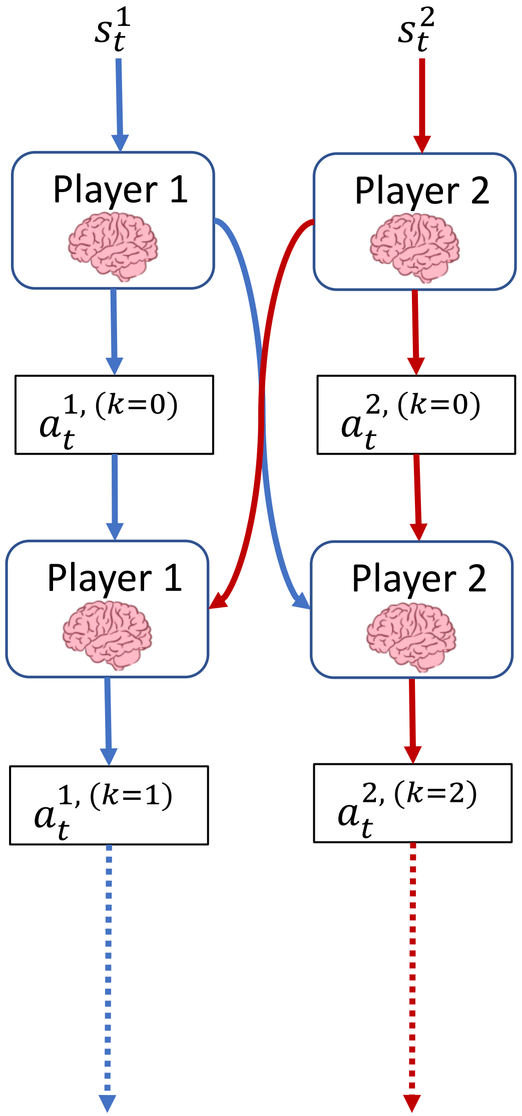
- Inspired by communication strategy in high-performing human teams, we propose iterated decision rationalization with mutual information for cooperative MARL



In this paper: Informational Policy Gradient (InfoPG)

- By assuming bounded-rational agents, we build a k-level, iterative architecture for InfoPG, inspired by the k-level reasoning from cognitive hierarchy theory[1].



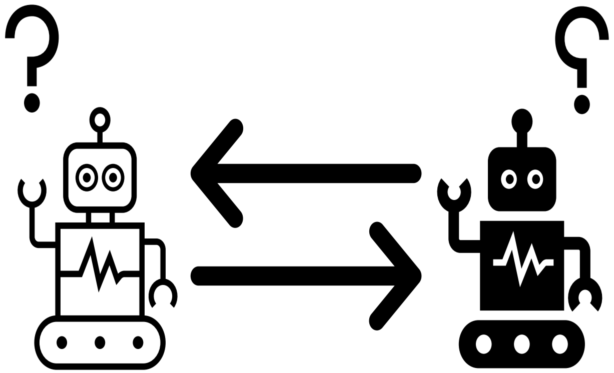


[1] Camerer, Colin F., Teck-Hua Ho, and Juin-Kuan Chong. "A cognitive hierarchy model of games." *The Quarterly Journal of Economics* 119.3 (2004): 861-898.

In this paper: Iterated decision rationalization with InfoPG for Cooperative MARL

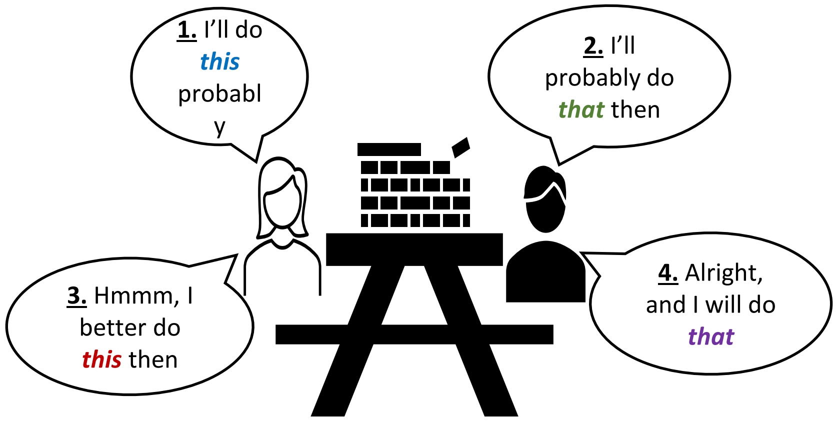
Basic Idea – Inspired by the k-level reasoning and assuming bounded rational agents:

- **We propose**, conditioning an agent's policy on its teammate's policies in a fully-decentralized setting



- **We hypothesize**, this conditionality inherently maximizes MI lower-bound among agents when optimizing under policy gradient

- **We hypothesize**, this maximization of MI lower-bound will improve MARL performance



Methodology: Informational Policy Gradient (InfoPG) Objective

- Pursuant to the general PG objective, we define the base form of the InfoPG objective as:

$$\nabla_{\theta}^{InfoPG} J(\theta) = \mathrm{E}_{\pi_{tot}^{i}} \left[G_t^i(o_t^i, a_t^i) \sum_{j \in \Delta_t^i} \nabla_\theta \log \left(\pi_{tot}^{i} \left(a_t^{i,\,(K)} \middle| a_t^{i,(k-1)}, a_t^{j\,(k-1)}, \ \ldots, o_t^i \right) \right) \right]$$

- Here $G_t^i(o_t^i, a_t^i)$ represents the return. We propose two variants of InfoPG where:

Implies non-negative reward from the env.

Only moves in the direction of maximizing MI

$$G_t^i(o_t^i, a_t^i) = Q_t^i(o_t^i, a_t^i) \quad \text{s.t.} \quad Q_t^i(o_t^i, a_t^i) \geq 0$$

Or

$$G_t^i(o_t^i, a_t^i) = A_t^i(o_t^i, a_t^i) = Q_t^i(o_t^i, a_t^i) - V_t^i(o_t^i)$$

InfoPG

Adv. InfoPG

Methodology: Connection to Mutual Information

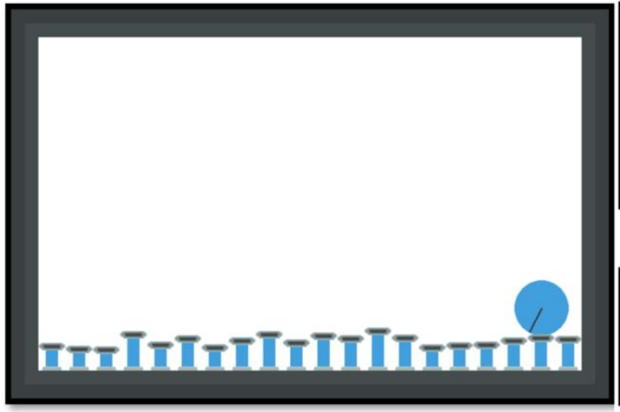
- MI is difficult to estimate in practice; but we derive a lower- and an upper-bound instead:

$$\boxed{\pi_{tot}^{i}(a^i|s^i,a^j)\log\left(\pi_{tot}^{i}(a^i|s^i,a^j)\right) \le I(\pi^i;\pi^j) \le 2\log(|A|) + 2\log\left(\pi_{tot}^{i}(a^i|s^i,a^j)\right)}$$

- Depending on the sign of $\nabla\pi_{tot}^{i}$, the bounds of $I(\pi^i;\pi^j)$ are "pushed" up or down
- In InfoPG with the non-negative reward condition always pushes up the MI lower-bound
- In Adv. InfoPG, the instantaneous sign of $\nabla\pi_{tot}^{i}$ depends on the sign of $A_t(o_t^i,a_t^i)$
  - If $A_t(o_t^i,a_t^i) > 0$ then the bounds of MI will shift ↑
  - If $A_t(o_t^i,a_t^i) < 0$ then the bounds of MI will shift ↓
- Over the full-extent of training Adv. InfoPG, MI is expected to increase as coordination improves

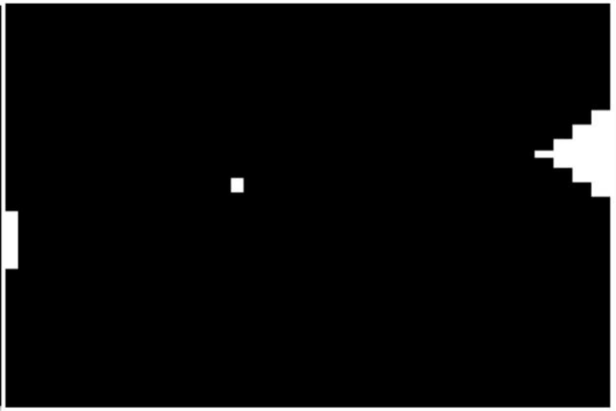
Adv. InfoPG modulates MI (rather than always maximizing it) depending on the cooperativity among agents and environment feedback.

Empirical Evaluation: Experiments and Evaluation Environments



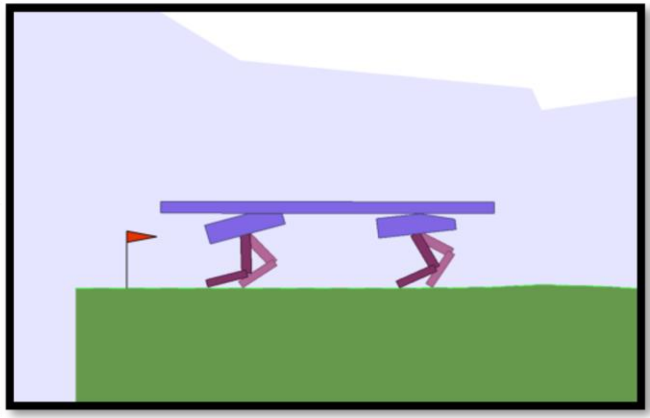
(a) Pistonball

Pistons work together to push a ball to the left wall by going up and down



(b) Co-op Pong

Paddles try to keep the ball in play for as long as possible by moving up and down



(c) Multiwalker

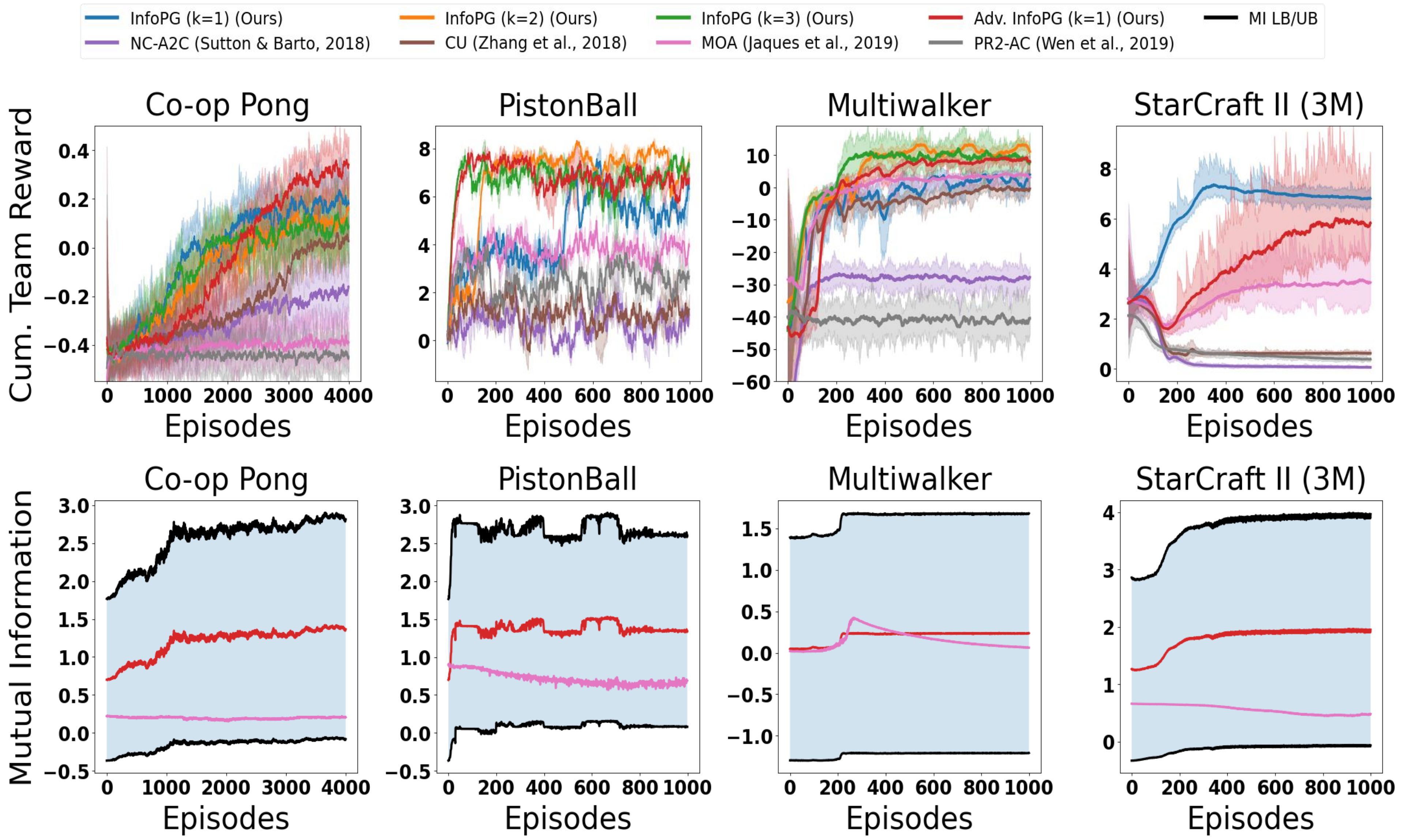
Bipedal walkers maintain individual balance and shared payload, while moving forward

(d) StarCraft II

Three Marines (allied) try to eliminate an enemy team of three Marines

Each of these games are decentralized, cooperative games.

Empirical Evaluation: Experimental Results (Training)



Summary

1- Adv. InfoPG wins Pong

2- Adv. InfoPG and InfoPG (k=2) win PistonBall

3- Adv. InfoPG and InfoPG (k=2, 3) win Multiwalker

4- InfoPG (k=1) wins StarCraft

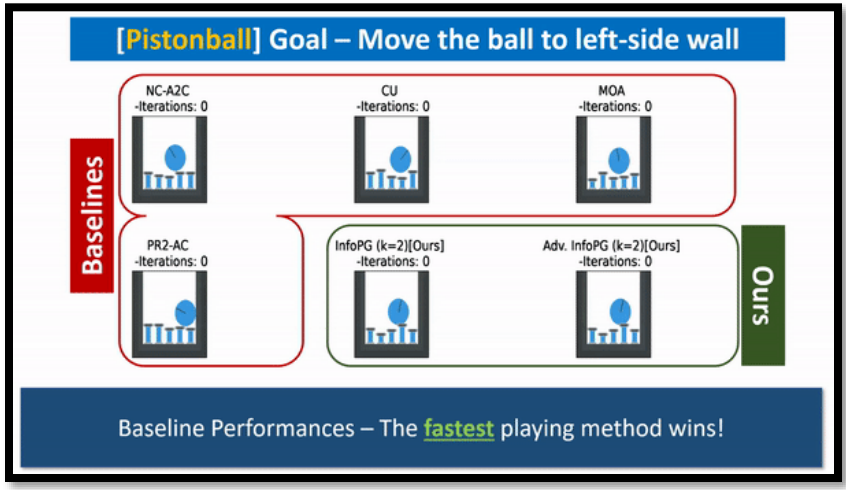
5- With Adv. InfoPG, MI increases over time with some instant fluctuations

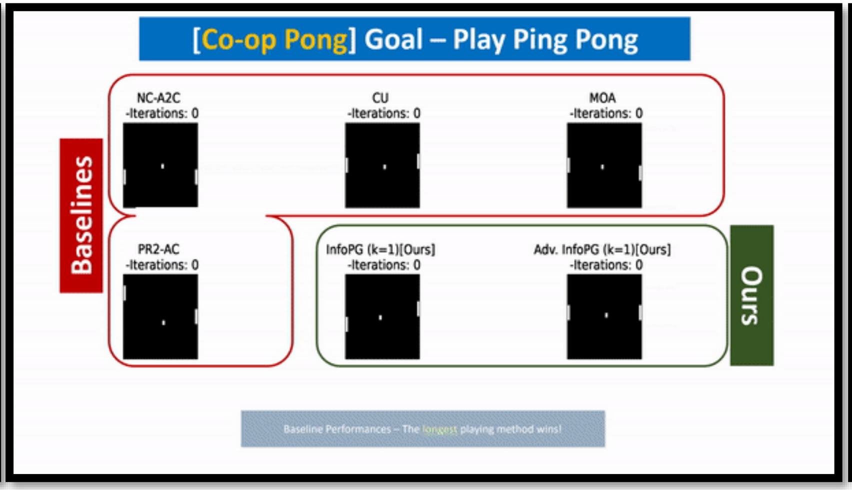
6- InfoPG MI > MOA MI over all experiments

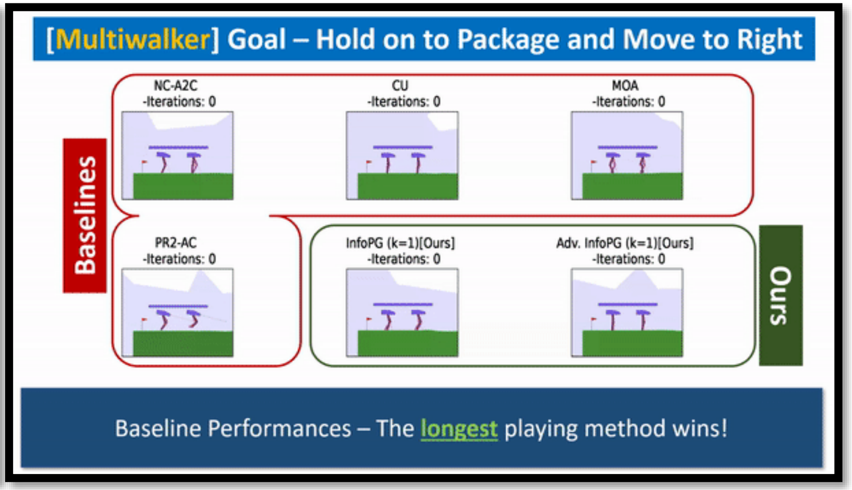
Empirical Evaluation: Experimental Results (Testing)

Table 1: Reported results are Mean (Standard Error) from 100 testing trials. For all tests, the final training policy at convergence is used for each method and for InfoPG and Adv. InfoPG, the best level of k is chosen.

Domain	**InfoPG**		**Adv. InfoPG**		MOA		CU		NC-A2C		PR2-AC	
	$\mathcal{R}$	#Steps	$\mathcal{R}$	#Steps	$\mathcal{R}$	#Steps	$\mathcal{R}$	#Steps	$\mathcal{R}$	#Steps	$\mathcal{R}$	#Steps
Co-op Pong	**0.17 (0.00)**	**203.2 (1.70)**	**0.22 (0.00)**	**213.6 (1.53)**	-0.4 (0.03)	44.7 (0.35)	0.05 (0.00)	134.7 (1.11)	-0.2 (0.00)	84.4 (0.93)	-0.85 (0.03)	38.7 (0.30)
Pistonball	**7.36 (0.02)**	**15.11 (0.22)**	**7.10 (0.02)**	**27.3 (0.40)**	3.73 (0.03)	82.6 (0.71)	0.89 (0.04)	146.6 (0.78)	0.86 (0.05)	141.9 (0.83)	-1.46 (0.04)	169 (0.71)
Multiwalker	**4.32 (0.10)**	**457.3 (1.08)**	**7.91 (0.08)**	**481.7 (0.80)**	4.21 (0.27)	460.8 (0.92)	1.852 (0.09)	179.6 (1.04)	-28 (0.12)	93.9 (0.43)	-155 (0.81)	147.3 (1.55)
StarCraft II	**6.47 (0.00)**	**30.1 (0.03)**	**5.40 (0.02)**	**43.5 (0.13)**	2.72 (0.01)	26.3 (0.05)	0.29 (0.00)	57.2 (0.09)	0.00 (0.00)	60.0 (0.00)	0.88 (0.04)	28.9 (0.09)



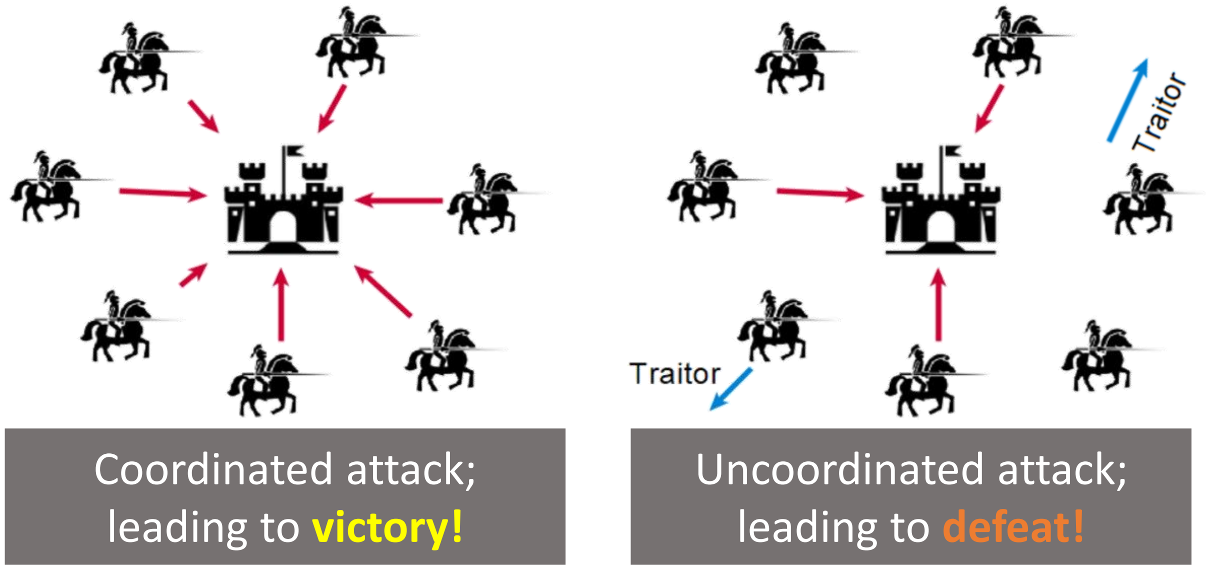




Video Link: https://youtu.be/rK_itCF9hPc

Empirical Evaluation : The Byzantine Generals Problem (BGP)

- We particularly studied Adv. InfoPG benefit by analyzing its performance in the Byzantine Generals Problem (BGP[2])
- The BGP describes a decision-making scenario in which involved agents must achieve consensus on an optimal collaborative strategy without relying on a trusted central party, but where at least one agent is corrupt and disseminates false information or is otherwise unreliable.

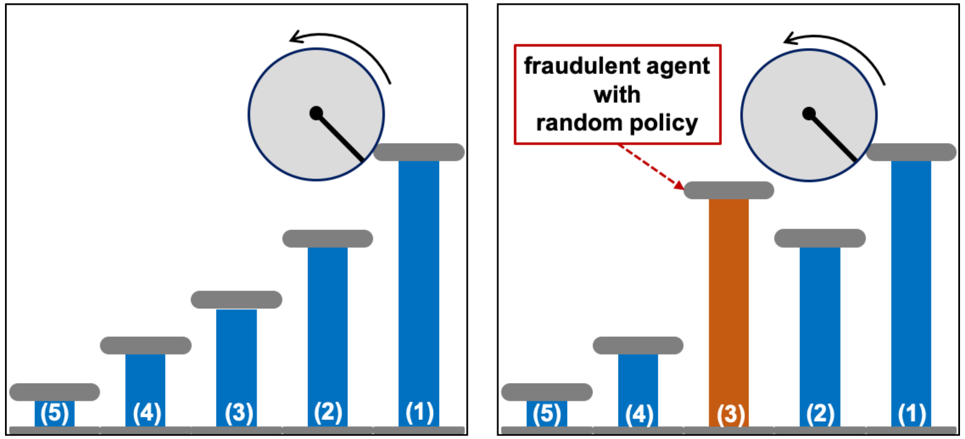


Curtesy of Medium. Available online at: https://medium.com/swlh/bitcoins-proof-of-work-the-problem-of-the-byzantine-generals-33dc4540442

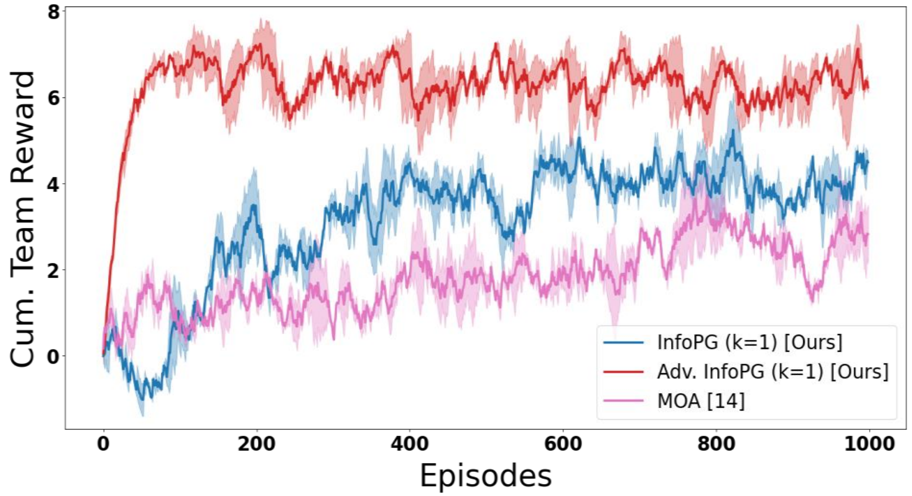
[2] Lamport, Leslie, Robert Shostak, and Marshall Pease. "The Byzantine generals problem." *Concurrency: the Works of Leslie Lamport*. 2019. 203-226.

Empirical Evaluation : The Byzantine Generals Problem (BGP)

- We particularly studied Adv. InfoPG benefit by analyzing its performance in the Byzantine Generals Problem (BGP[2])
- The BGP describes a decision-making scenario in which involved agents must achieve consensus on an optimal collaborative strategy without relying on a trusted central party, but where at least one agent is corrupt and disseminates false information or is otherwise unreliable.
- We designed a BGP scenario in Pistonball where there is one ”faulty” agent who the other agents shouldn’t listen to



(numbers represent piston agent indices)



Summary

Adv. InfoPG attains larger cumulative rewards because agents learn not to maximize mutual information with Piston #3

[2] Lamport, Leslie, Robert Shostak, and Marshall Pease. "The Byzantine generals problem." *Concurrency: the Works of Leslie Lamport*. 2019. 203-226.

Conclusions

- InfoPG is a framework for decentralized, cooperative MARL and implicit MI maximization without the need for auxiliary regularization terms.
- InfoPG uses a k-level theory of mind to deeply rationalize agents' action-decisions.
- InfoPG sets a new SOTA against other decentralized baselines in learning emergent cooperative policies in complex, discrete/continuous domains.
- Results between InfoPG and Adv. InfoPG, as well in the BGP scenario show that strict-non-negative MI maximization may not always be desirable.
- Adv. InfoPG modulates MI among agents, rather than always maximizing it, to improve coordination based on agents' observed cooperativity and environment feedback.

Questions?



Full-Read: https://arxiv.org/pdf/2201.08484.pdf

Paper **Demo** **Code**

Thank you!