Iterated Reasoning with Mutual Information in Cooperative and Byzantine Decentralized Teaming

Sachin Konan*, Esmaeil Seraj*, and Matthew Gombolay
*Co-first authors (these authors contributed equally to this work). *Corresponding author: Esmaeil Seraj* <email: eseraj3@gatech.edu>

Introduction: Information Sharing for Multi-Agent Teaming

- A team usually entails a group of individuals who have a shared, common objective requiring the team to take actions according to the mutual interest(s) of the group.
- Information sharing, or communication, is a key feature in building team cognition.

Communication → Coordination → Collaboration

- Much like us humans or some animal species, robots also need to communicate in order to coordinate their actions and perform as a team.

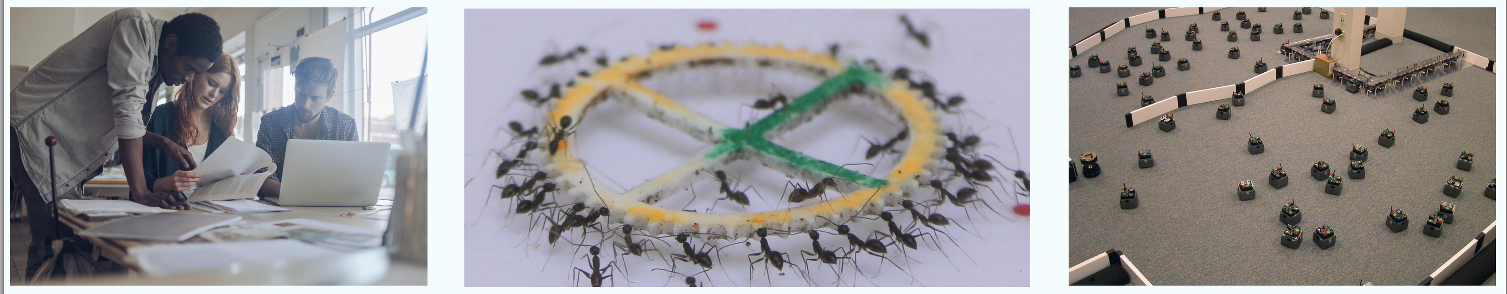


Figure 1. Examples of human, animal, and robot teams.

Motivation: Iterated Communication and Rationalizability

- High-performing human teams, not only use communication, but they also benefit from acting strategically with hierarchical levels of iterated communication and rationalizability.

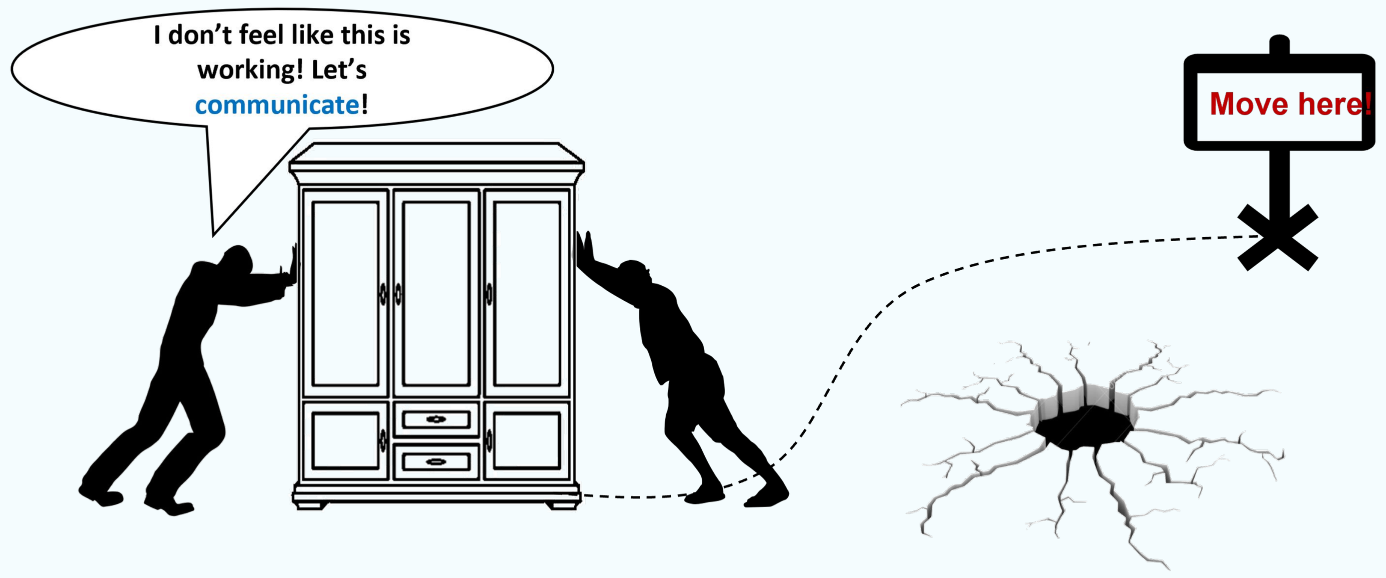


Figure 2. it is too strong to assume all teammates are perfectly rational in their decision-making.

- Yet, most of the prior work in MARL does not support iterated decision-rationalization and by assuming perfectly informed, rational agents, only encourage inter-agent communication, resulting in a suboptimal equilibrium cooperation strategy.
- Inspired by communication strategy in high-performing human teams, we propose **InfoPG** which leverages iterated decision-rationalization with mutual information for cooperative MARL.

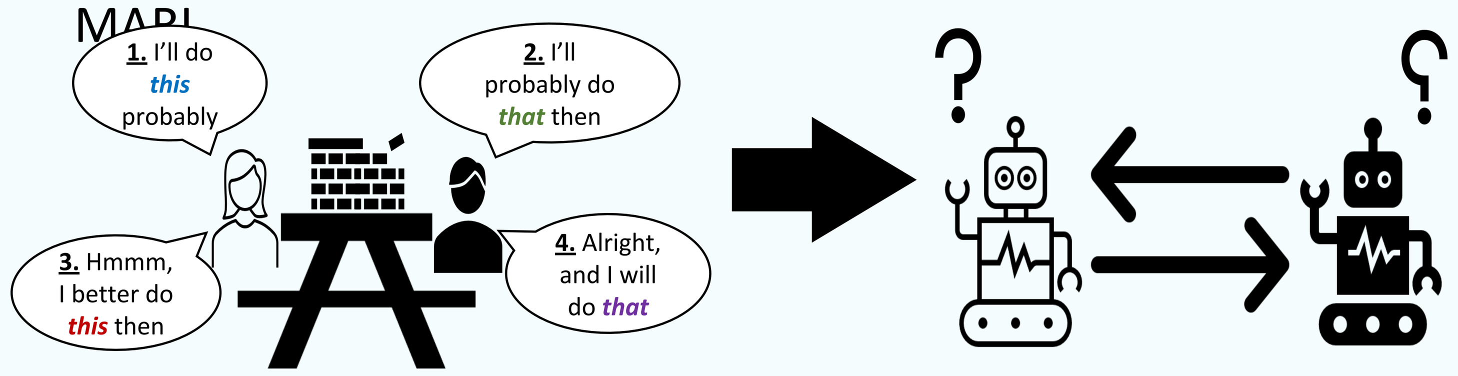


Figure 3. InfoPG enables iterated reasoning for robot decision-making.

Informational Policy Gradient (InfoPG): Algorithmic Overview and Big-Picture

- By assuming bounded-rational agents, we build a k-level, iterative architecture for **InfoPG**, inspired by the k-level reasoning from cognitive hierarchy theory.
- In **InfoPG**, each agent is equipped with an **encoding** and a **communicative** policy.

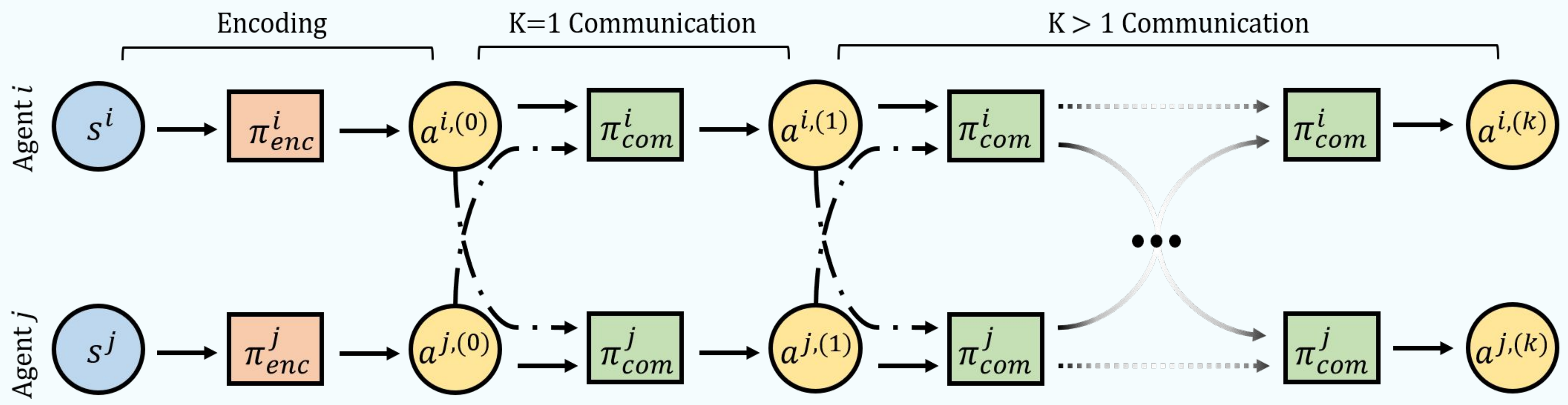


Figure 4. Decision rationalization in InfoPG.

- **InfoPG steps**: Depending on the decision-rationalization depth k…
  1. At the beginning of a new rollout, each agent receives a state observation and produces an initial action $\pi^i_{enc}(a^{i,(0)}|\, o^i)$
  2. Agents locally communicate their action guesses as high-dimensional latent distributions with neighboring agents
  3. Agents repeat step #2 k times and update their action-guesses iteratively using their communicative policy $\pi^i_{comm}(a^{i,(k)}|a^{i,(k-1)}, a^{j,(k-1)}, \ldots, o^i)$

InfoPG Variants: Objective Function and Connection to Mutual Information (MI)

- Pursuant to the general Policy Gradient objective, we define the **InfoPG** objective as:

$$\nabla_\theta^{InfoPG} J(\theta) = \mathrm{E}_{\pi^i_{tot}}\left[G^i_t(o^i_t, a^i_t) \sum_{j\in\Delta^i_t} \nabla_\theta \log\left(\pi^i_{tot}\left(a^{i,(K)}_t \middle| a^{i,(k-1)}_t, a^{j\,(k-1)}_t, \ldots, o^i_t\right)\right)\right]$$

- Here $G^i_t(o^i_t, a^i_t)$ represents the return. We propose two variants of **InfoPG** where:

$$G^i_t(o^i_t, a^i_t) = Q^i_t(o^i_t, a^i_t) \quad \text{s.t.} \quad Q^i_t(o^i_t, a^i_t) \geq 0 \quad \text{InfoPG}$$

Or

$$G^i_t(o^i_t, a^i_t) = A^i_t(o^i_t, a^i_t) = Q^i_t(o^i_t, a^i_t) - V^i_t(o^i_t) \quad \text{Adv. InfoPG}$$

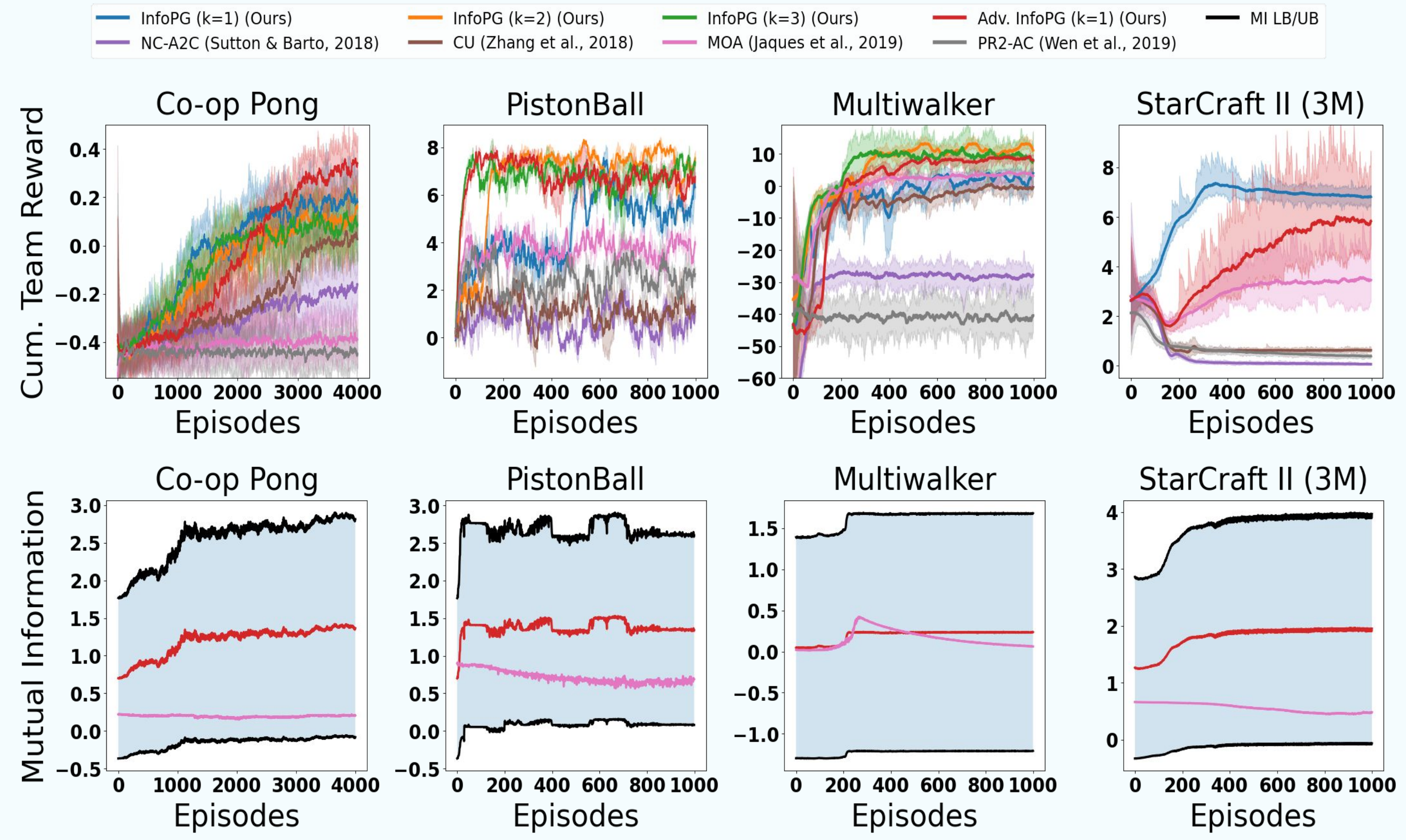
- We derive a lower- and an upper-bound on the MI between agents' policy distributions:

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

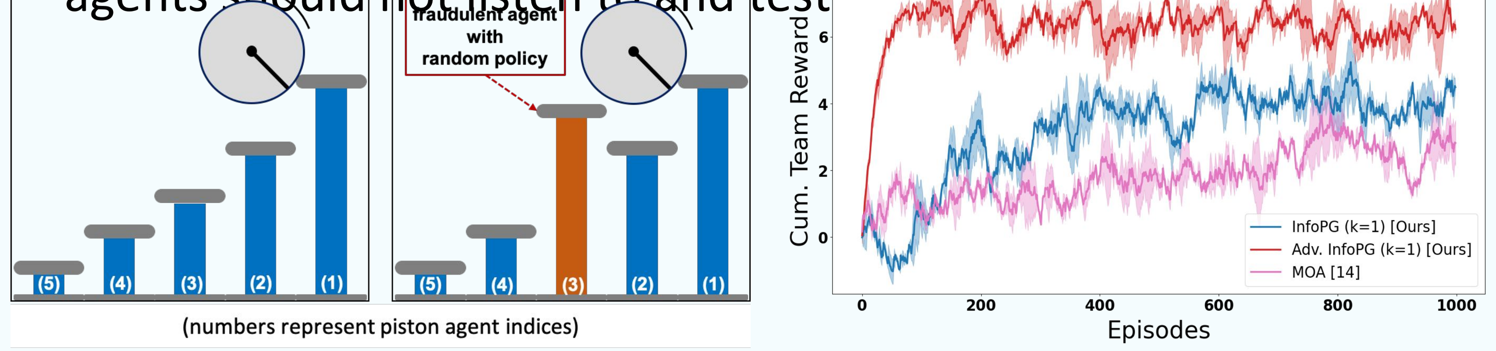
- Depending on the sign of $\nabla\pi^i_{tot}$, the bounds of $I(\pi^i;\pi^j)$ are "pushed" up or down
- In **InfoPG** with the non-negative reward condition always pushes up the MI lower-bound
- In **Adv. InfoPG**, the instantaneous sign of $\nabla\pi^i_{tot}$ depends on the sign of $A_t(o^i_t, a^i_t)$
  - If $A_t(o^i_t, a^i_t) > 0$ then the bounds of MI will shift ↑
  - If $A_t(o^i_t, a^i_t) < 0$ then the bounds of MI will shift ↓
- **Adv. InfoPG** modulates MI (rather than always maximizing it) depending on the cooperativity among agents and environment feedback.

Empirical Evaluation: Results & Experiments

- We benchmark **InfoPG** and **Adv. InfoPG** against **NC-A2C**, **CU**, **MOA**, and **PR2-AC** in four fully-decentralized, cooperative domains: Pistonball, Co-op Pong, Multiwalker, and StarCraft II.



- **In summary**, we show that not only **InfoPG** and **Adv. InfoPG** achieve higher cumulative results and better sample-efficiency than the baseline methods, but they also resulted in higher MI among agents, leading to a higher quality action coordination.
- **The Byzantine Generals Problem (BGP) Scenario:** The BGP describes a 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 with untrainable random policy who the other agents should not listen to and test



- **In summary**, **Adv. InfoPG** attains larger cumulative rewards as agents learn not to maximize mutual information with Piston #3

Conclusions

- **InfoPG** tackles decentralized, cooperative MARL with implicit MI maximization, which uses a k-level theory of mind to deeply rationalize agents' action-decisions.
- Results of **InfoPG** and **Adv. InfoPG**, in the **BGP scenario** show that always maximizing the MI may not always be desirable.

- Demo:
- Code:
- Full-read: