Learning to Negotiate via Voluntary Commitment

Shuhui Zhu

University of Waterloo & Vector Institute
shuhui.zhu@uwaterloo.ca

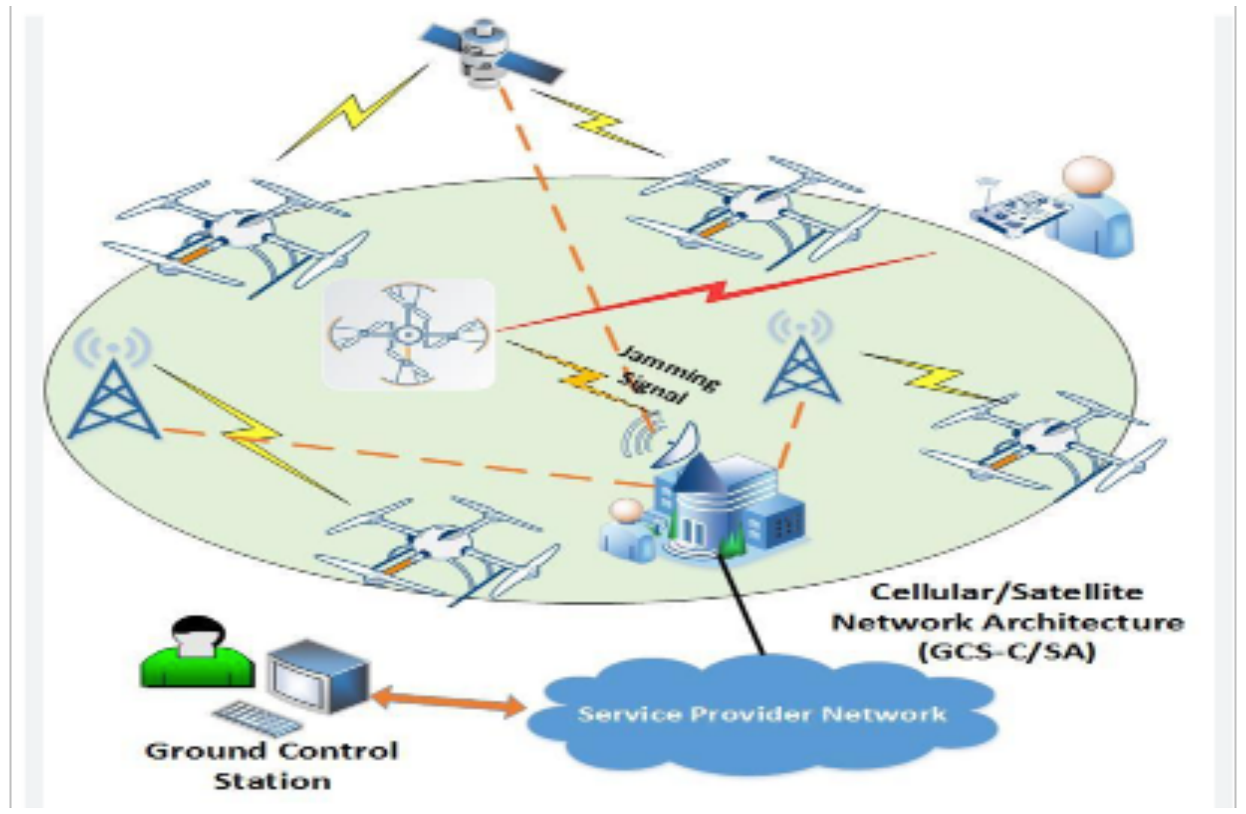
Why Cooperative AI Matters?

Why Cooperative AI Matters: Enhancing Intelligent Multi-Agent Systems

As AI systems become more widely deployed, **they will inevitably interact with each other** across a broader range of domains.

Autonomous Driving

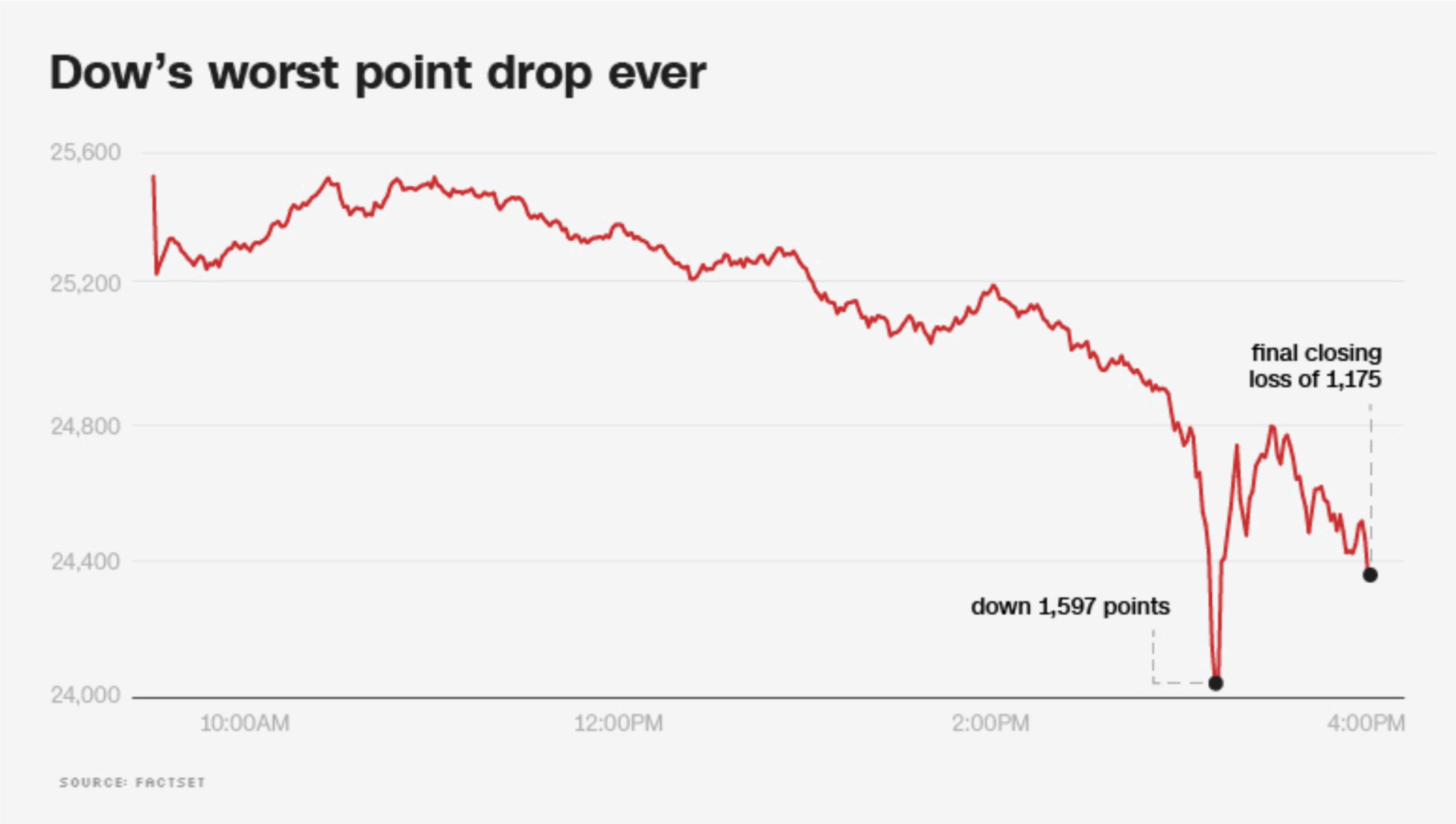
E-Commerce



UAV Surveillance

Smart Grids

Why Cooperative AI Matters: Avoiding Disastrous Outcomes

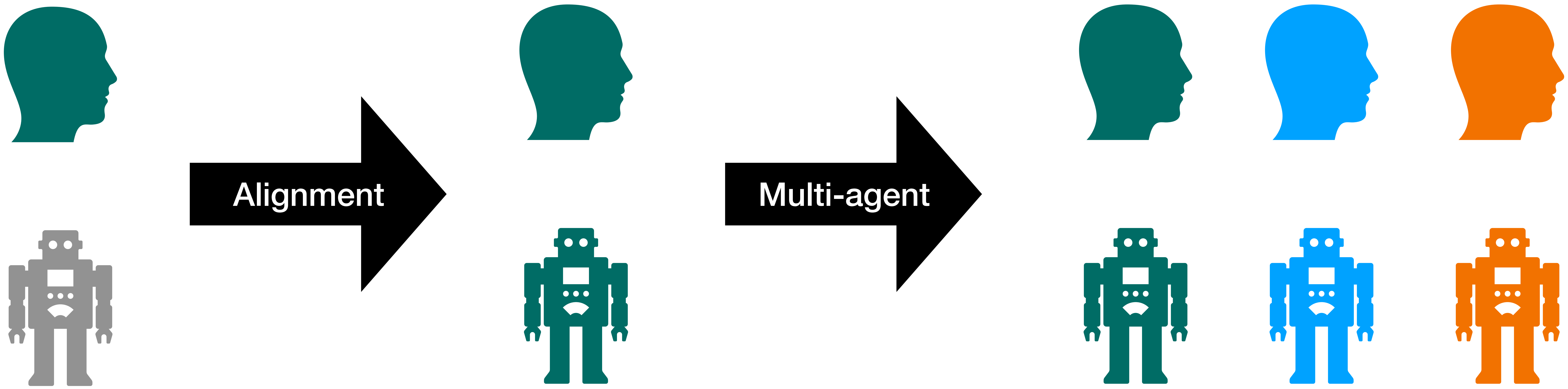


The 2010 Flash Crash

A bizarre domino effect triggered by **high-frequency trading (HFT) algorithms** erased almost **1 trillion** in market value.

Cooperation Problems

Cooperation Problems



Even if each agent individually is well aligned with human values, they may **fail to cooperate** due to **mixed interests.**

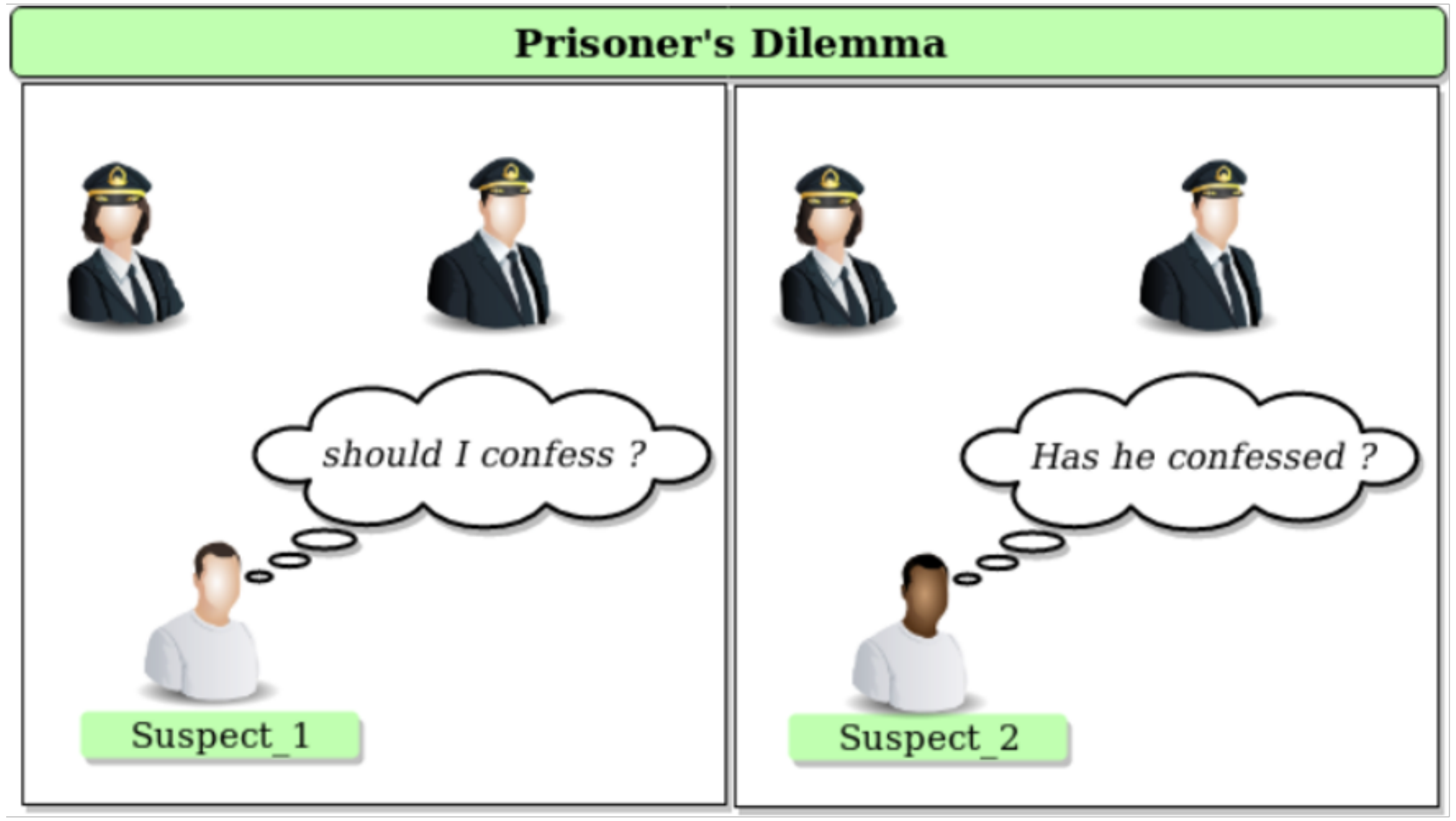
Understanding and communication alleviate cooperation problems in low conflict level

Fully Cooperative Games

HUNTER 2 \ HUNTER 1	STAG	HARE
STAG	5, 5	0, 2
HARE	2, 0	1, 1

Trust Dilemma

Challenges of Communication in Highly Conflicting Games



	C	D
C	(-1,-1)	(-3,0)
D	(0,-3)	(-2,-2)

Table 1: Prisoner's dilemma

Regardless of the opponent's statements or actions, each rational prisoner will choose to **defect**.

What If They Can Make Conditional Commitments?

Conditional Commitments in Iterated Prisoner's Dilemma

Grim Trigger

I will cooperate as long as you do. However, if you defect even once, I will **permanently switch to defection.**

Tit-for-Tat

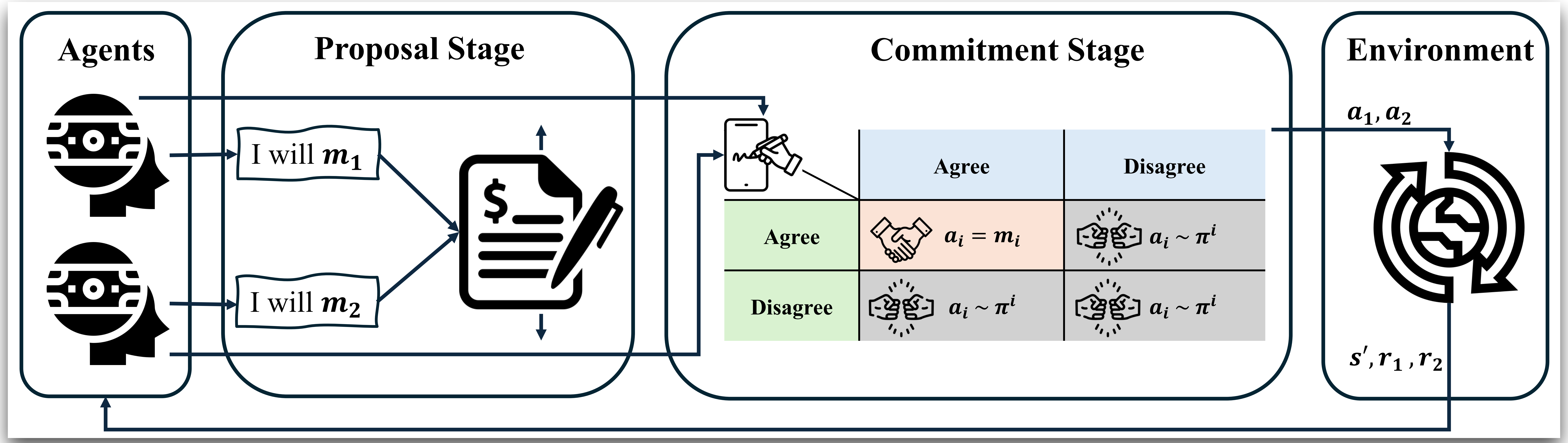
I will begin by cooperating and will always **mirror your last action.**

How to Design Smart Adpative Commitments?

Methodology

Markov Commitment Games

Outline



Markov Commitment Games

Notation

$$MCG = (\mathcal{N}, \mathcal{S}, \mathcal{T}, (\mathcal{M}^i, \mathcal{C}^i, \mathcal{A}^i, \mathcal{R}^i)_{i \in \mathcal{N}}, \gamma)\,.$$

- $\mathcal{N}$: The set of agents (players) in the game, indexed by $i \in \mathcal{N}$.
- $\mathcal{S}$: The state space, representing all possible states of the environment.
- $\mathcal{M}^i$: The proposal space of agent i.
- $\mathcal{C}^i$: The commitment space of agent i.
- $\mathcal{A}^i$: The action space of agent i, the joint action space is $\mathcal{A} = (\mathcal{A}^i)_{i \in \mathcal{N}}$.
- $\mathcal{R}^i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$: The reward function of agent i.
- $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$: The environment transition function, which satisfies the Markov property and the stationarity condition, i.e., $\mathcal{T}(s_{t+1} \,|\, s_t, \mathbf{a}_t) = \mathcal{T}(s_{t+1} \,|\, s_t, \mathbf{a}_t, \ldots, s_0, \mathbf{a}_0) = \mathcal{T}(s' \,|\, s, \mathbf{a})$.

If we consider only a single agent's action in a multi-agent environment, the environment can become **non-stationary** from that agent's perspective.

Markov Commitment Games

Notation

$$MCG = (\mathcal{N}, \mathcal{S}, \mathcal{T}, (\mathcal{M}^i, \mathcal{C}^i, \mathcal{A}^i, \mathcal{R}^i)_{i \in \mathcal{N}}, \gamma) \, .$$

In an MCG, each agent i has three decisions to make at each time step:

- Proposal policy $\phi^i_{\eta^i} : \mathcal{S} \to \Delta(\mathcal{M}^i)$.
- Commitment policy $\psi^i_{\zeta^i} : \mathcal{S} \times \mathcal{M} \to \Delta(\mathcal{C}^i)$.
- Action policy, $\pi^i_{\theta^i} : \mathcal{S} \to \Delta(\mathcal{A}^i)$.

Mutual Cooperation Becomes an Equilibrium in Prisoner’s Dilemma

Proposition 4.1.

Mutual cooperation is a ***Pareto-dominant Nash equilibrium*** *in the MCG of the Prisoner's Dilemma.*

	C	D
C	(-1,-1)	(-3,0)
D	(0,-3)	(-2,-2)

Table 1: Prisoner’s dilemma

I will propose cooperation.
I will commit to a joint proposal where my coplayer proposes cooperation, and reject otherwise.
I will choose defection if there is no mutual agreement.

How to **Learn** Smart Adpative Commitments?

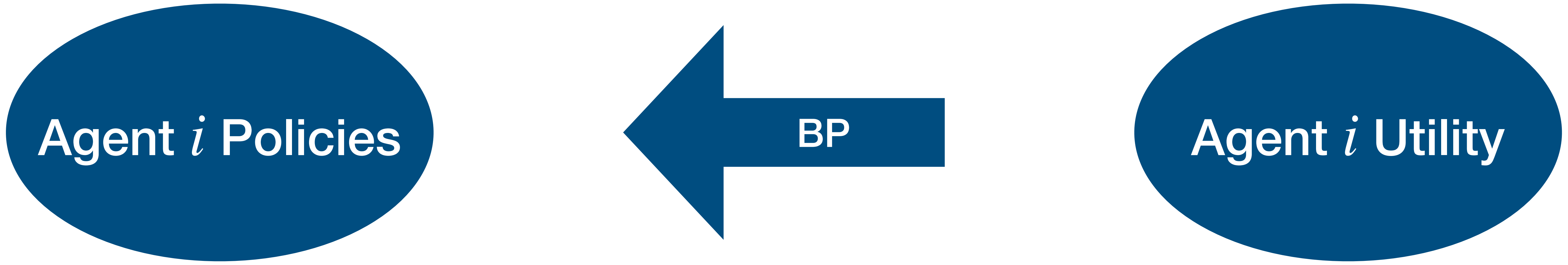
Differentiable Commitment Learning

Objective

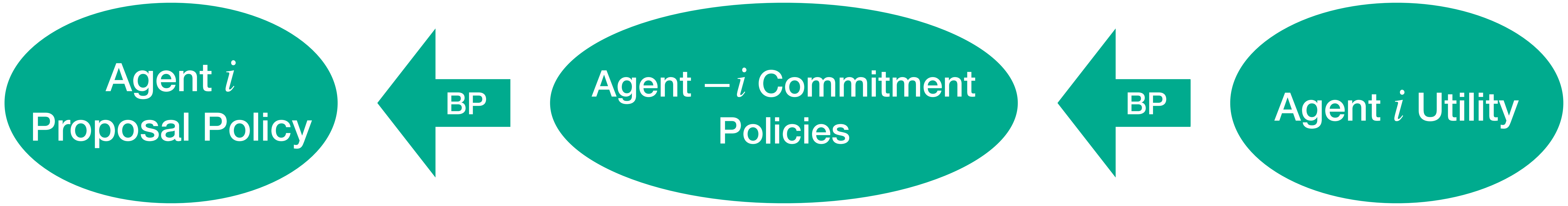
$$\max_{\eta^i,\zeta^i,\theta^i} V^i_{\phi,\psi,\pi}(s) = \mathbb{E}_{\phi,\psi,\pi}\left[\sum_{k=t}^{\infty} \gamma^{k-t} r^i_{k+1} \mid s_t = s\right]$$

Environment dynamics are influenced by all agents' policies

Direct Effect



Indirect Effect



Lemma 5.1.

Given proposal policy $\phi^i_{\eta^i}$, commitment policy $\psi^i_{\zeta^i}$ and the action policy $\pi^i_{\theta^i}$ of each agent i in an MCG, the gradients of the value function $V^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(s)$ w.r.t. θ^i, ζ^i, η^i are

$$\nabla_{\theta^i} V^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(s) \propto \mathbb{E}_{x\sim\rho_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}},\mathbf{m}\sim\boldsymbol{\phi},\mathbf{c}\sim\boldsymbol{\psi},\mathbf{a}\sim\boldsymbol{\pi}}\Big[\Big(1-\mathbf{1}(\mathbf{c}=\mathbf{1})\Big)Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(x,\mathbf{a})\,\nabla_{\theta^i}\log\pi^i(a^i\,|\,x)\Big],$$

$$\begin{aligned}\nabla_{\zeta^i} V^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(s) \propto \mathbb{E}_{x\sim\rho_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}},\mathbf{m}\sim\boldsymbol{\phi},\mathbf{c}\sim\boldsymbol{\psi},\mathbf{a}\sim\boldsymbol{\pi}}\Bigg[&\Big[\mathbf{1}(\mathbf{c}=\mathbf{1})Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(x,\mathbf{m})+\Big(1-\mathbf{1}(\mathbf{c}=\mathbf{1})\Big)Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(x,\mathbf{a})\Big]\,\nabla_{\zeta^i}\log\psi^i(c^i\,|\,x,\mathbf{m})\\&+\Big[Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(x,\mathbf{m})-Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(x,\mathbf{a})\Big]\prod_{k\neq i}\mathbf{1}(c^k=1)\cdot\nabla_{\zeta^i}\mathbf{1}(c^i=1)\Bigg],\end{aligned}$$

$$\begin{aligned}\nabla_{\eta^i} V^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(s) \propto \mathbb{E}_{x\sim\rho_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}},\mathbf{m}\sim\boldsymbol{\phi},\mathbf{c}\sim\boldsymbol{\psi},\mathbf{a}\sim\boldsymbol{\pi}}\Bigg[&\Big[\mathbf{1}(\mathbf{c}=\mathbf{1})Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(x,\mathbf{m})+\Big(1-\mathbf{1}(\mathbf{c}=\mathbf{1})\Big)Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(x,\mathbf{a})\Big]\cdot\Big(\nabla_{\eta^i}\log\phi^i(m^i\,|\,x)+\sum_j\nabla_{\eta^i}\log\psi^j(c^j\,|\,x,\mathbf{m})\Big)\\&+\sum_j\prod_{k\neq j}\mathbf{1}(c^k=1)\Big[Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(x,\mathbf{m})-Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(x,\mathbf{a})\Big]\cdot\nabla_{\eta^i}\mathbf{1}(c^j=1)\Bigg],\end{aligned}$$

where $Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(s,\mathbf{a})=\mathbb{E}_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}[\sum_{k=t}^{\infty}\gamma^{k-t}r^i_{k+1}\,|\,s_t=s,\mathbf{a}_t=\mathbf{a}]\,.$

Incentive-Compatible Constraints Encourage Mutually Beneficial Proposals

Agents may still have the **equilibrium selection problem** when multiple equilibria exist.

Incentive-Compatible Constraints

$$\mathbb{E}_{\mathbf{m}\sim\phi}[Q^i_{\phi,\psi,\pi}(s,\mathbf{m})] \geq \mathbb{E}_{\mathbf{a}\sim\pi}[Q^i_{\phi,\psi,\pi}(s,\mathbf{a})] \quad \forall\ i\ .$$

Mutually Beneficial Deals Do Not Exist

$\phi^i(s) = \pi^i(s), \forall i$

$\mathbb{E}_{\mathbf{m}\sim\phi}[Q^i_{\phi,\psi,\pi}(s,\mathbf{m})] = \mathbb{E}_{\mathbf{a}\sim\pi_U}[Q^i_{\phi,\psi,\pi}(s,\mathbf{a})], \forall i$

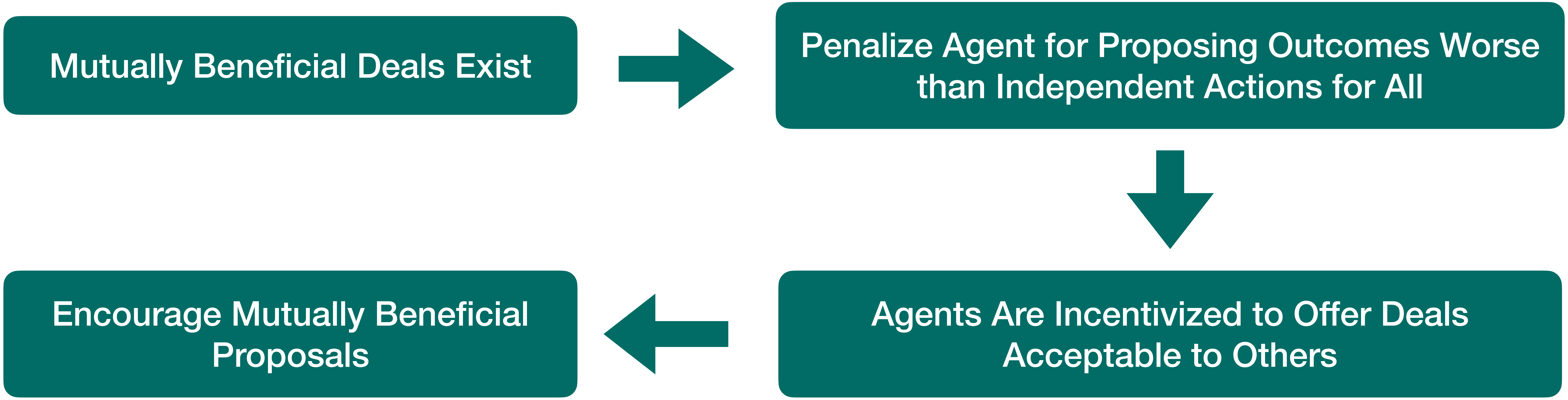
Feasible Solutions Always Exist

Incentive-Compatible Constraints Encourage Mutually Beneficial Proposals

Agents may still have the **equilibrium selection problem** when multiple equilibria exist.

Incentive-Compatible Constraints

$$\mathbb{E}_{\mathbf{m}\sim\phi}[Q^i_{\phi,\psi,\pi}(s,\mathbf{m})] \geq \mathbb{E}_{\mathbf{a}\sim\pi}[Q^i_{\phi,\psi,\pi}(s,\mathbf{a})] \quad \forall\ i$$



Integrate Incentive-Compatible Constraints into the Objective

$$\eta^i \leftarrow \eta^i + \nabla_{\eta^i} V^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(s) + \lambda \nabla_{\eta^i} \sum_j \min\{0, \mathbb{E}_{\mathbf{m}\sim\boldsymbol{\phi}}[Q^j_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(s, \mathbf{m})] - \mathbb{E}_{\mathbf{a}\sim\boldsymbol{\pi}}[Q^j_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(s, \mathbf{a})]\}.$$

Improve expected self-return

Increase the likelihood that its proposals are accepted by others

- The incentive-compatible constraints are **applied to the proposal policy only**.
- If a proposal is acceptable to others but does not benefit the ego agent, the **commitment policy is trained to reject non-profitable proposals**, reinforcing self-interest.

Empirical Results

Evaluated Methods

Centralized DCL: have full access to others' **actual** policies and critics.

- DCL: $\lambda = 0$.
- DCL-IC: $\lambda = 1$.

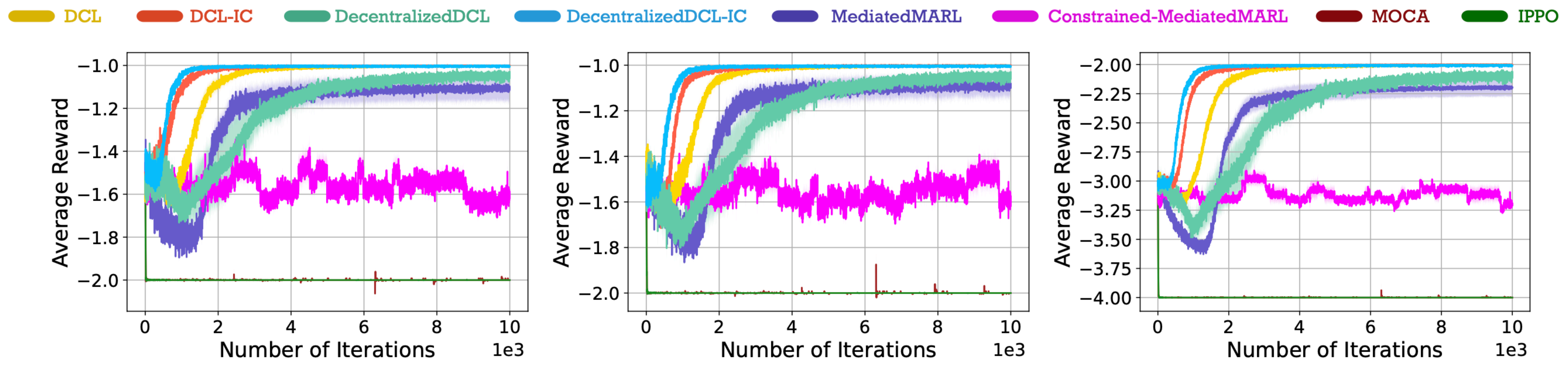
Decentralized DCL: need to **estimate** others' actual policies and critics.

- DecentralizedDCL: $\lambda = 0$.
- DecentralizedDCL-IC: $\lambda = 1$.

IPPO: each agent was trained independently with the proximal policy optimization (PPO).

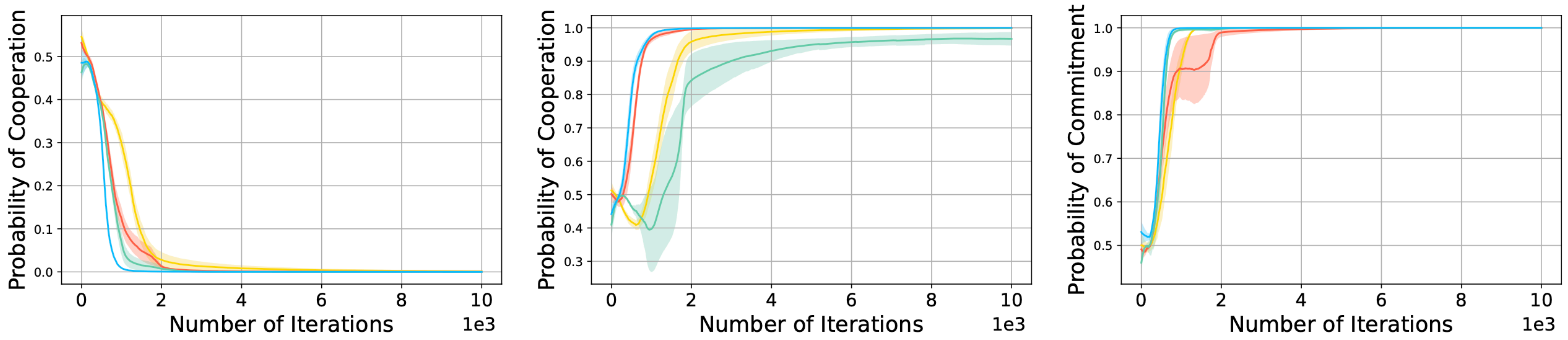
Mediated-MARL: altruistic joint planner was trained to maximize the utilitarian social welfare.

MOCA: Each agent was trained to maximize self-interest, with a learnable transfer payment that directly modifies agents' rewards.



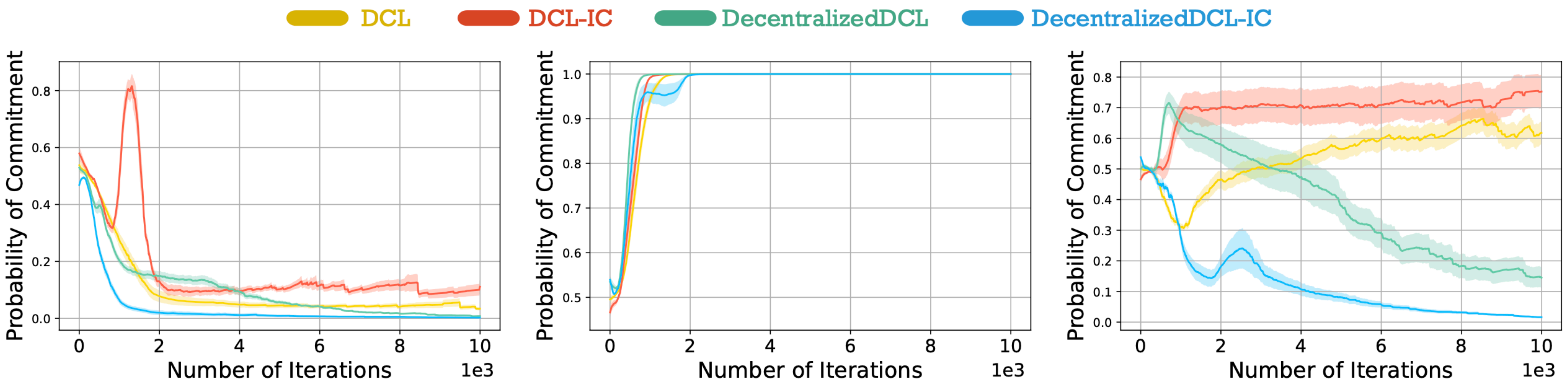
(a) Agent 1 Reward in PD (b) Agent 2 Reward in PD (c) Social Welfare in PD

Figure 2: Prisoner's Dilemma: DCL v.s. Other Baselines



(a) DCL Action Policy (b) DCL Proposal Policy (c) DCL Commitment Policy of (C, C)

Figure 3: DCL Policies in Prisoner's Dilemma



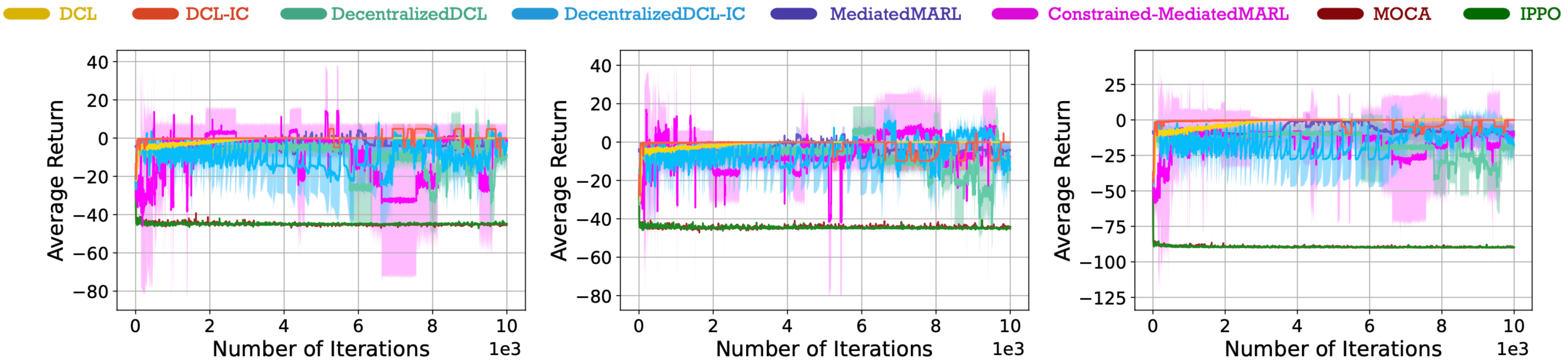
(a) Commitment Policy of (C, D) (b) Commitment Policy of (D, C) (c) Commitment Policy of (D, D)

Figure 6: DCL Commitment Policies in Prisoner's Dilemma

DCL agents strategically accept beneficial agreements while rejecting disadvantageous ones.

Resilient against malicious agents who always propose defection.

Sequential Social Dilemma



(a) Agent 1 Return in Grid Game (b) Agent 2 Return in Grid Game (c) Social Welfare in Grid Game

Figure 4: Grid Game (Horizon=16): DCL v.s. Other Baselines.

Repeated Purely Conflicting Game

Table 2: Purely Conflicting Game

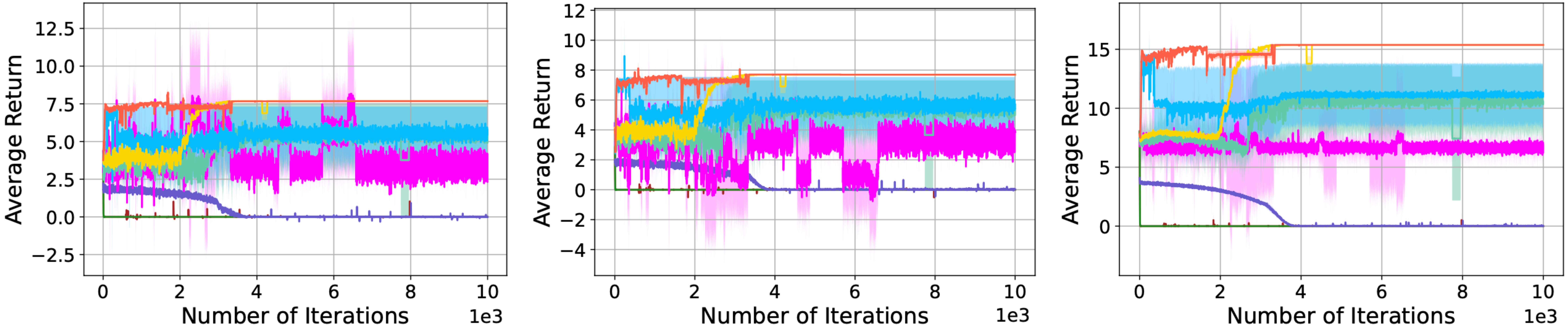
	A_1	A_2
A_1	(0,0)	(-1,2)
A_2	(2,-1)	(0,0)

Agents cannot establish **one-step** mutually beneficial agreements.

If agents can commit to actions over **multiple steps**, both can achieve positive long-term returns by committing to a **tit-for-tat** agreement.

extended DCL with **mega-step commitments**

Repeated Purely Conflicting Game



(a) Agent 1 Return in RPC (b) Agent 2 Return in RPC (c) Social Welfare in RPC

Figure 5: Repeated Purely Conflicting Game (Horizon=16): DCL v.s. Other Baselines.

Take-home Messages

- Agents can achieve mutually beneficial outcomes by voluntarily committing to proposed actions in **MCGs**.
- **DCL** enables agents to learn strategic commitments by differentiating through self policies (direct effect) and others' policies (indirect effect).
- **Incentive-compatible learning** accelerates agreement formation by encouraging agents to propose agreements that will be accepted by others.
- **Mega-step commitments** can enhance long-term cooperation in some repeated purely competitive environments.