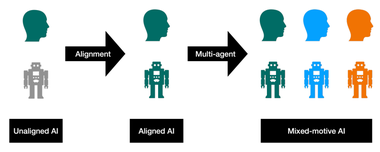
Learning to Negotiate via Voluntary Commitment

Shuhui Zhu[1,2], Baoxiang Wang[3], Sriram Ganapathi Subramanian[2], Pascal Poupart[1,2]
[1]University of Waterloo, [2]Vector Institute, [3]The Chinese University of Hong Kong, Shenzhen

shuhui-zhu.github.io

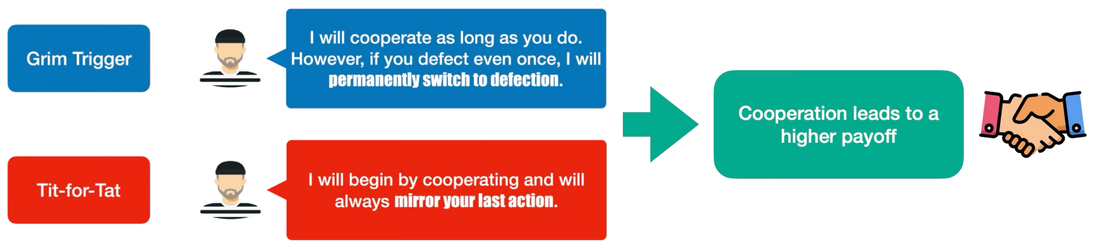
Introduction

Cooperation Problems: The partial alignment and conflict of autonomous agents lead to **mixed-motive** scenarios in many real-world applications. However, even if each agent individually is well aligned with human values, they may fail to cooperate due to mixed interests, even when cooperation yields a better outcome.



	Cooperate	Defect
Cooperate	(-1, -1)	(-3, 0)
Defect	(0, -3)	(-2, -2)

Commitment Mechanism: If agents can make **conditional commitment** based on other agents' policies, cooperation may become an equilibrium even in highly conflicting environments.



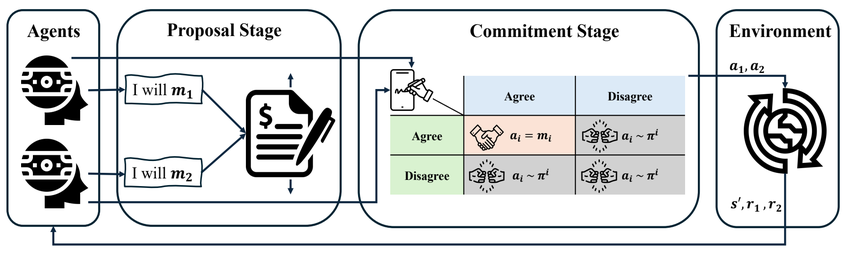
Objective: We aim to develop a **learnable** commitment protocol that allows **self-interested** agents to strategically align their actions, thereby **enhancing cooperation in mixed-motive multi-agent systems**. This approach also enhances the **accuracy** of value estimations and promotes **stability** during training. We empirically showed that our method outperforms the baseline methods in multiple tasks.

Markov Commitment Games Framework

We introduce the Markov Commitment Games, a framework that allows self-interested agents to negotiate future plans through voluntary commitments.

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

- $\mathcal{N}$: 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., $\mathcal{T}(s_{t+1}\,|\,s_t, \mathbf{a}_t) = \mathcal{T}(s_{t+1}\,|\,s_t, \mathbf{a}_t, \ldots, s_0, \mathbf{a}_0) = \mathcal{T}(s'|\,s, \mathbf{a})$.
- γ: The discount factor.



In an MCG, each agent i has three decisions to make at each time step, with the objective of maximizing its expected cumulative return

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

- 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, $\pi^i_{\theta^i} : \mathcal{S} \to \Delta(\mathcal{A}^i)$.

Proposition 4.1.

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

Differentiable Commitment Learning Algorithm

Under the framework of MCGs, we propose differentiable commitment learning (DCL), which **maximizes agents' expected self-interests** while incorporating **incentive-compatible constraints** on their proposal policies to encourage mutually beneficial agreements.

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 $V^i_{\phi,\psi,\pi}(s)$ w.r.t. θ^i, ζ^i, η^i are

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

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

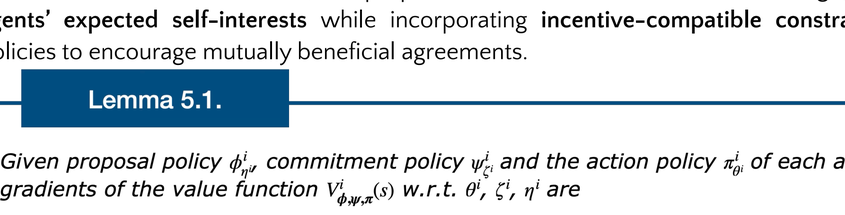
$$\nabla_{\eta^i} V^i_{\phi,\psi,\pi}(s) \propto \mathbb{E}_{x\sim\rho_{\phi,\psi,\pi},\mathbf{m}\sim\phi,\mathbf{c}\sim\psi,\mathbf{a}\sim\pi}\Bigg[\Big[\mathbf{1}(\mathbf{c}=\mathbf{1})Q^i_{\phi,\psi,\pi}(x,\mathbf{m}) + \Big(1-\mathbf{1}(\mathbf{c}=\mathbf{1})\Big)Q^i_{\phi,\psi,\pi}(x,\mathbf{a})\Big]\cdot\Big(\nabla_{\eta^i}\log\phi^i(m^i\,|\,x)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

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

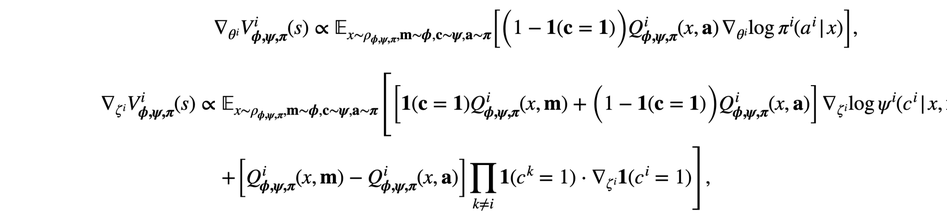
DCL enables agents to optimize strategies by considering both **direct and indirect effects** on their utilities.

- Direct impact: Agents **differentiate through their own policies** to maximize individual returns.
- Indirect impact: Agents anticipate how their decisions affect others' commitments and, in turn, their own outcomes—captured via **differentiation through others' commitment policies**.

Code Link



We further propose incentive-compatible learning to encourage agents to find mutually beneficial proposals.



Experiments

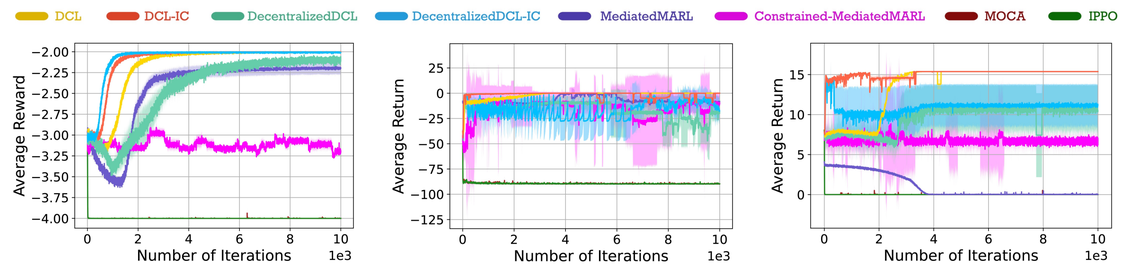
Evaluated Methods

- **DCL** (Ours): Centralized and decentralized variants, with/without incentive-compatible constraints.
- **Independent PPO**: Agents trained independently using PPO.
- **Mediated-MARL**: Altruistic planner trained to optimize utilitarian social welfare.
- **MOCA**: Self-interested agents with learnable transfer payments modifying rewards.

Evaluation Scenarios

- **Prisoner's Dilemma**: Tabular social dilemma.
- **Grid Game**: Sequential social dilemma.
- **Repeated Purely Conflicting Game**: Iterated setting with strictly opposing interests.

Results



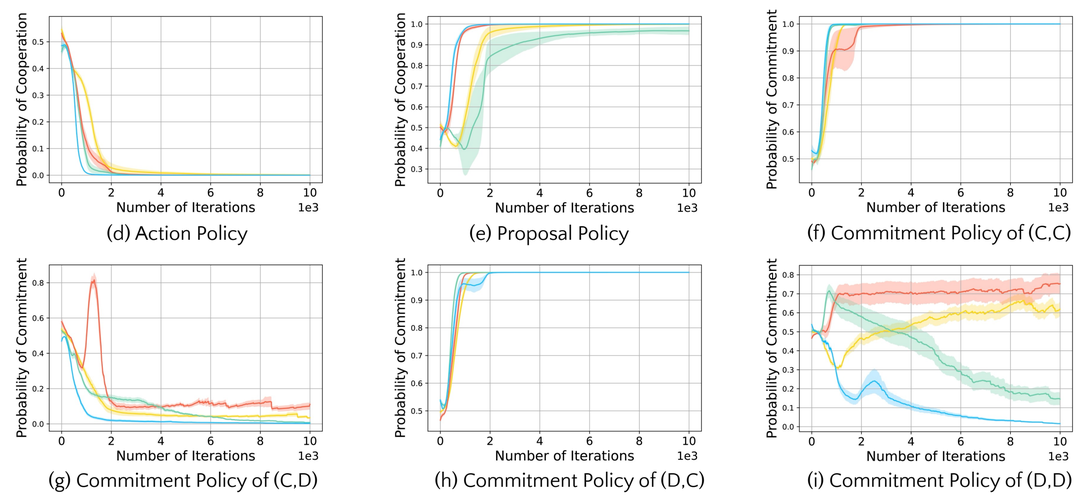
(a) Social Welfare in PD (b) Social Welfare in Grid Game (c) Social Welfare in RPC

Highlights of Results

- Both centralized and decentralized DCL **outperform all baselines** across the three scenarios.
- Incentive-compatible DCL **converges faster** than unconstrained DCL.
- In the repeated pure conflict game, **multi-step commitments** enable agents to **sustain positive returns** through tit-for-tat agreements.

Examples of Learned Policies (Prisoner's Dilemma)

- Without mutual commitment, DCL agents converge to mutual defection.
- With conditional commitment, agents propose and commit to mutual cooperation (probability → 1).
- DCL agents learn to **accept mutually beneficial agreements** and **reject proposals that exploit them** (e.g., co-player defects while they propose cooperation).



(d) Action Policy (e) Proposal Policy (f) Commitment Policy of (C,C)

(g) Commitment Policy of (C,D) (h) Commitment Policy of (D,C) (i) Commitment Policy of (D,D)

Discussion

Scalability to Many-player Scenarios

In MCGs, the joint proposal space grows exponentially with the number of agents, which would inevitably increase the computational complexity. To investigate how DCL handles scalability with many players, we conducted additional experiments on an N-player public goods game.

- Most agents converge to propose contributions and commit to joint proposals that result in positive individual welfare.
- These findings indicate that DCL scales well to many-player games, with the agreement rate of joint proposals remaining stable (>0.99) as the number of agents increases.

Number of Agents	Run Time (Hours)	Agreement Rate	Social Welfare
2	4	0.996 ± 0.002	0.997 ± 0.002
3	7	0.994 ± 0.001	1.491 ± 0.004
5	12	0.996 ± 0.001	1.989 ± 0.002
10	32	0.991 ± 0.001	3.659 ± 0.143

Robustness to Maliciously Irrational Agents

- **Robustness**: DCL agents learn to accept mutually beneficial commitments and reject exploitative ones.
- **Adaptive Behavior**: When paired with irrational or malicious agents who always propose defection, DCL agents refuse harmful proposals.
- **Outcome**: This behavior demonstrates DCL's robustness against malicious behavior by safeguarding self-interest through strategic commitment.

Conclusion

- Agents **achieve mutually beneficial outcomes by voluntarily committing** to proposed actions in MCGs.
- DCL enables strategic commitment learning by **differentiating through both self and others' policies**.
- **Incentive-compatible learning** accelerates agreement formation by encouraging agents to propose agreements that will be accepted by others.
- **Mega-step commitments** enhance long-term cooperation in repeated competitive settings.

Limitations and Future Work

Sample Efficiency

- Limitation: On-policy updates are less sample efficient and challenging in decentralized settings.
- Future Work: Explore efficient, bias-robust training methods for decentralized multi-agent learning.

Complex Proposal Domain

- Limitation: Current proposals are limited to deterministic future actions.
- Future Work: Extend the framework to support stochastic or conditional commitments.

Acknowledgements

We acknowledge funding from the Canada CIFAR AI Chair program, a discovery grant from the Natural Sciences and Engineering Research Council of Canada and a grant from IITP & MSIT of Korea. Computational resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute.