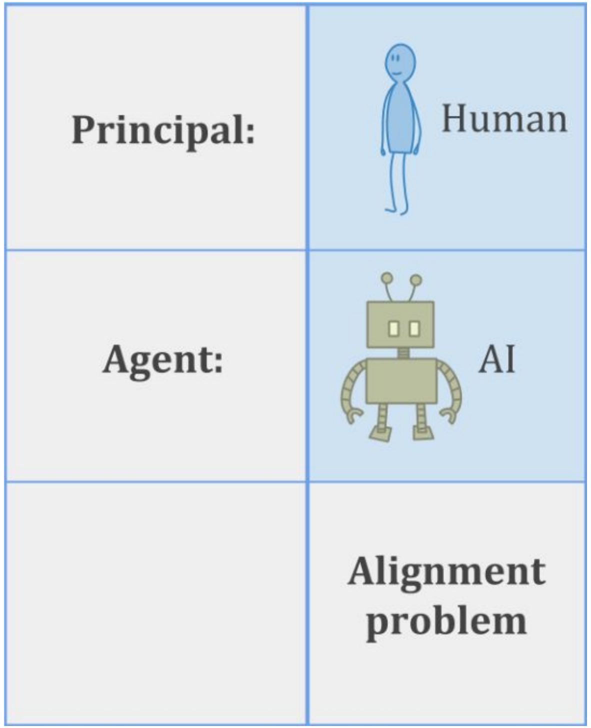
Evaluating LLMs in Open-Source Games

Swadesh Sistla, Max Kleiman-Weiner

UNIVERSITY of WASHINGTON

Humanity may increasingly delegate agency to AI

Core Multi-Agent Challenges



Principal: Human

Agent: AI

Alignment problem

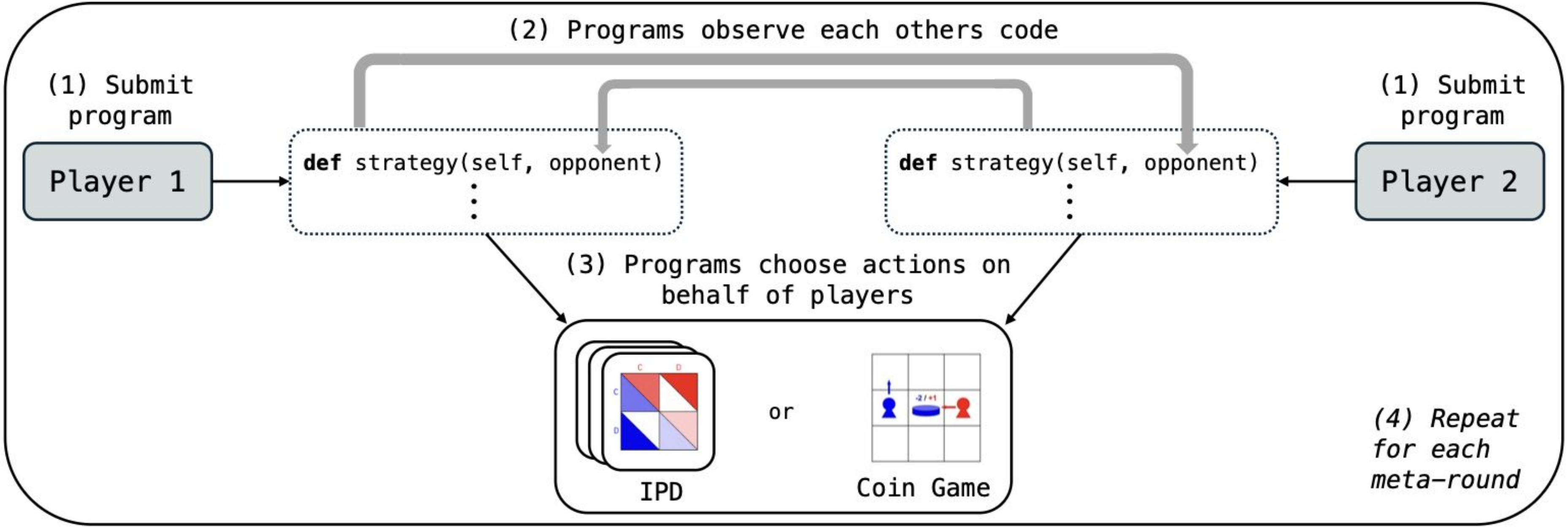
Multi-agent Alignment

		Player B	
		Cooperate	Defect
Player A	Cooperate	(3, 3)	(0, 5)
	Defect	(5, 0)	(2, 2)

Multi-agent Cooperation

How can we build AIs that intelligently cooperate?

One Idea: Open-Source Game Theory



(2) Programs observe each others code

(1) Submit program

Player 1

```
def strategy(self, opponent)
                :
```

```
def strategy(self, opponent)
                :
```

(1) Submit program

Player 2

(3) Programs choose actions on behalf of players

IPD

or

Coin Game

(4) Repeat for each meta–round

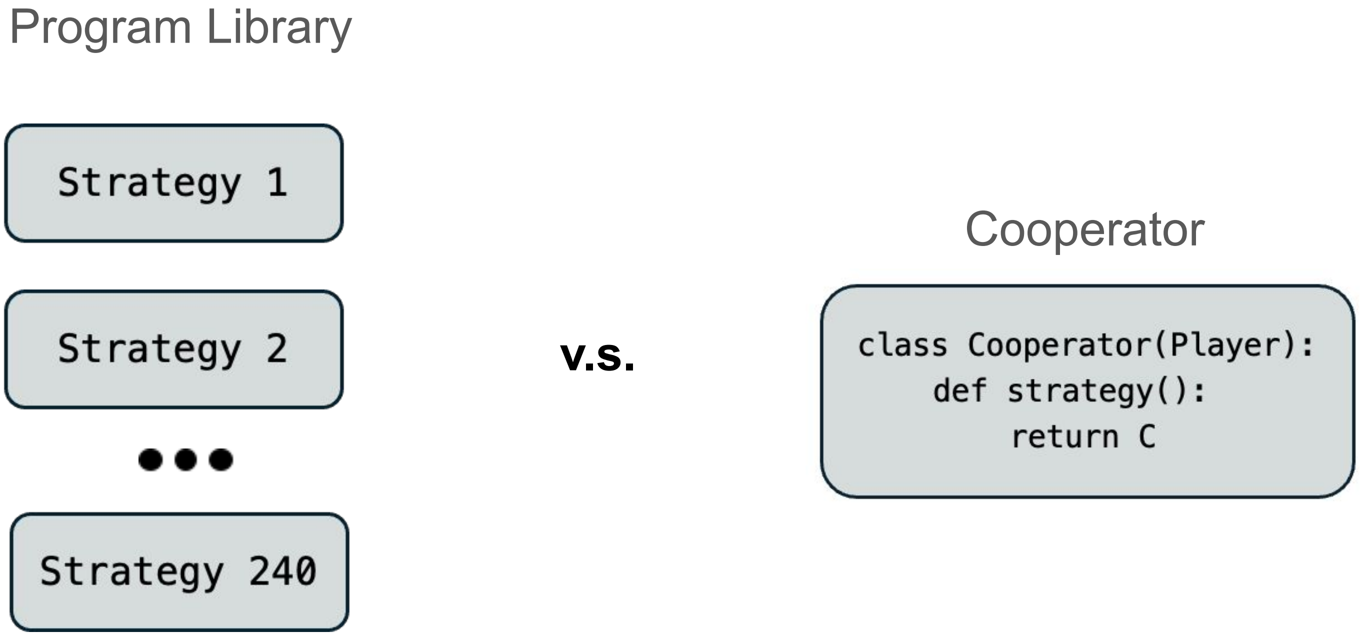
Submit actions → Submit <u>programs</u>

Question: Can today’s LLMs participate in open-source games? If so, what kind of behavior emerges?

Answer: See our paper!

First: Can AIs reason about strategic code?

SPARC Benchmark: Predicting Reciprocal Cooperation



Program Library

Strategy 1

Strategy 2

•••

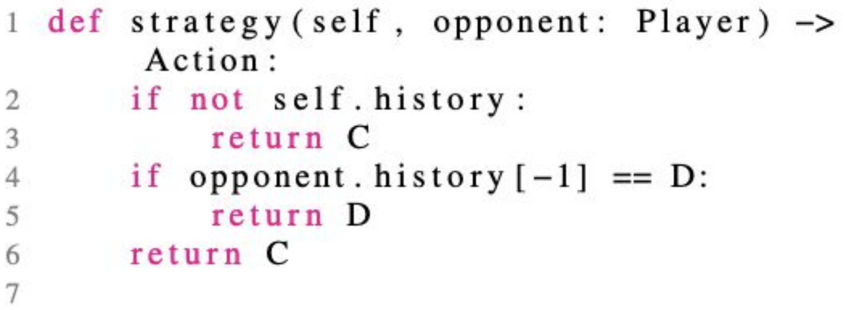
Strategy 240

v.s.

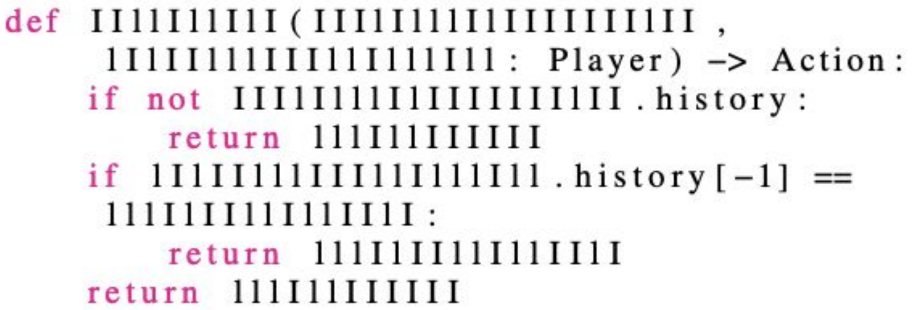
Cooperator

```
class Cooperator(Player):
    def strategy():
        return C
```

Isolating Strategic Reasoning



```
def strategy(self, opponent: Player) ->
    Action:
    if not self.history:
        return C
    if opponent.history[-1] == D:
        return D
    return C
```

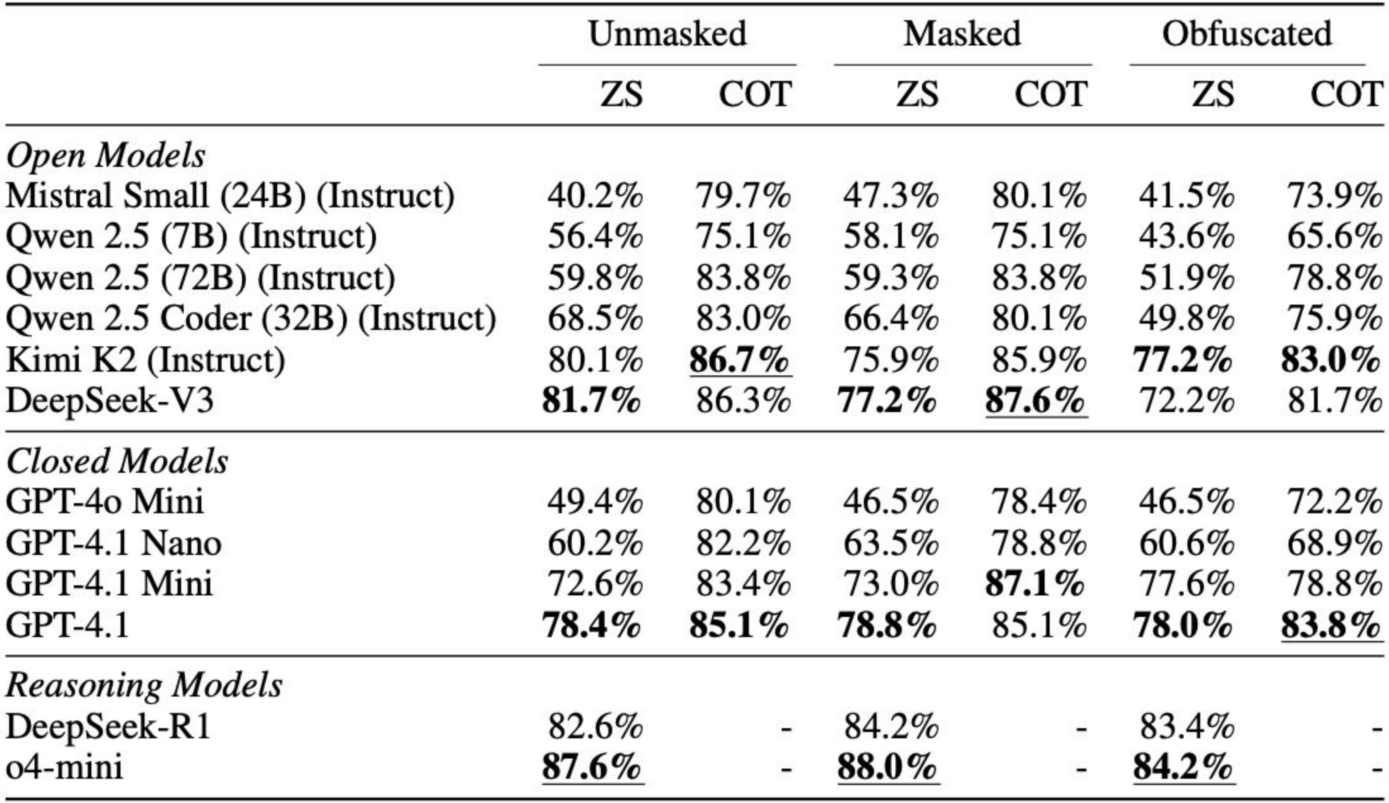


```
def IIllIllIlI(IIIlIllIlIIIIIIIlII,
     lIlIIlllIIIllIlllIll: Player) -> Action:
    if not IIIlIllIlIIIIIIIlII.history:
        return lllIllIIIII
    if lIlIIlllIIIllIlllIll.history[-1] ==
     lllIlIIllIllIIlI:
        return lllIlIIllIllIIlI
    return lllIllIIIII
```

Masking

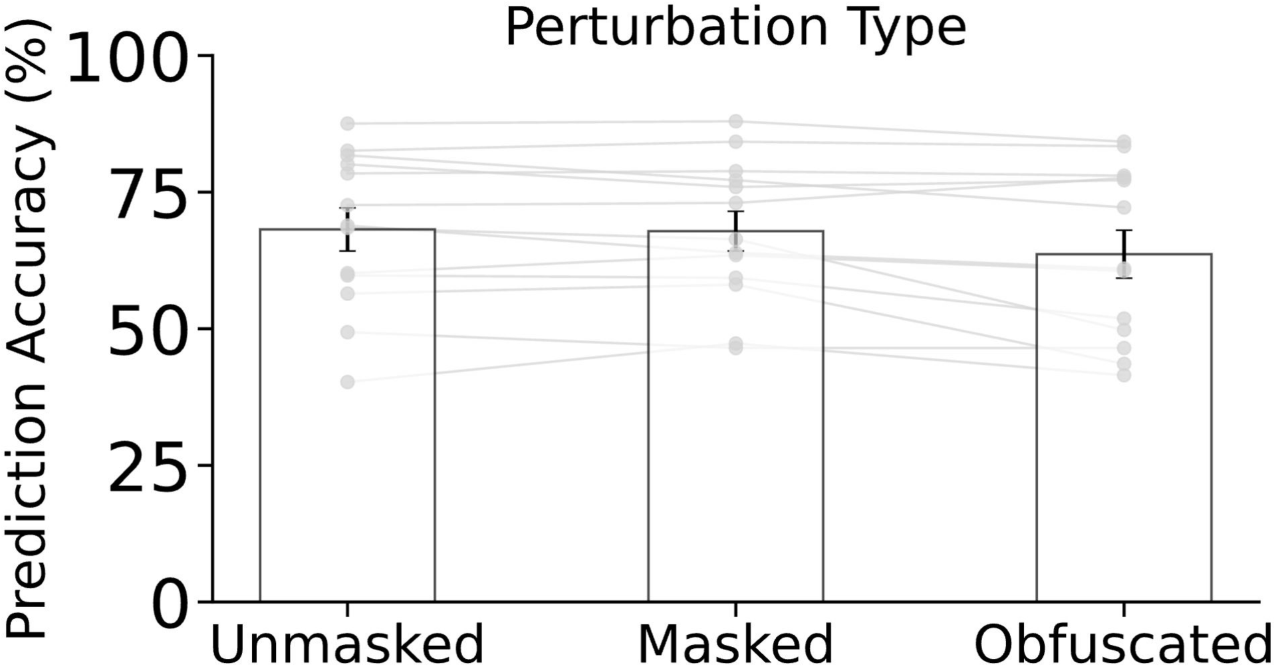
Obfuscation

Classification Results (%)



	Unmasked		Masked		Obfuscated	
	ZS	COT	ZS	COT	ZS	COT
Open Models						
Mistral Small (24B) (Instruct)	40.2%	79.7%	47.3%	80.1%	41.5%	73.9%
Qwen 2.5 (7B) (Instruct)	56.4%	75.1%	58.1%	75.1%	43.6%	65.6%
Qwen 2.5 (72B) (Instruct)	59.8%	83.8%	59.3%	83.8%	51.9%	78.8%
Qwen 2.5 Coder (32B) (Instruct)	68.5%	83.0%	66.4%	80.1%	49.8%	75.9%
Kimi K2 (Instruct)	80.1%	**86.7%**	75.9%	85.9%	**77.2%**	**83.0%**
DeepSeek-V3	**81.7%**	86.3%	**77.2%**	**87.6%**	72.2%	81.7%
Closed Models						
GPT-4o Mini	49.4%	80.1%	46.5%	78.4%	46.5%	72.2%
GPT-4.1 Nano	60.2%	82.2%	63.5%	78.8%	60.6%	68.9%
GPT-4.1 Mini	72.6%	83.4%	73.0%	**87.1%**	77.6%	78.8%
GPT-4.1	**78.4%**	**85.1%**	**78.8%**	85.1%	**78.0%**	**83.8%**
Reasoning Models						
DeepSeek-R1	82.6%	-	84.2%	-	83.4%	-
o4-mini	**87.6%**	-	**88.0%**	-	**84.2%**	-

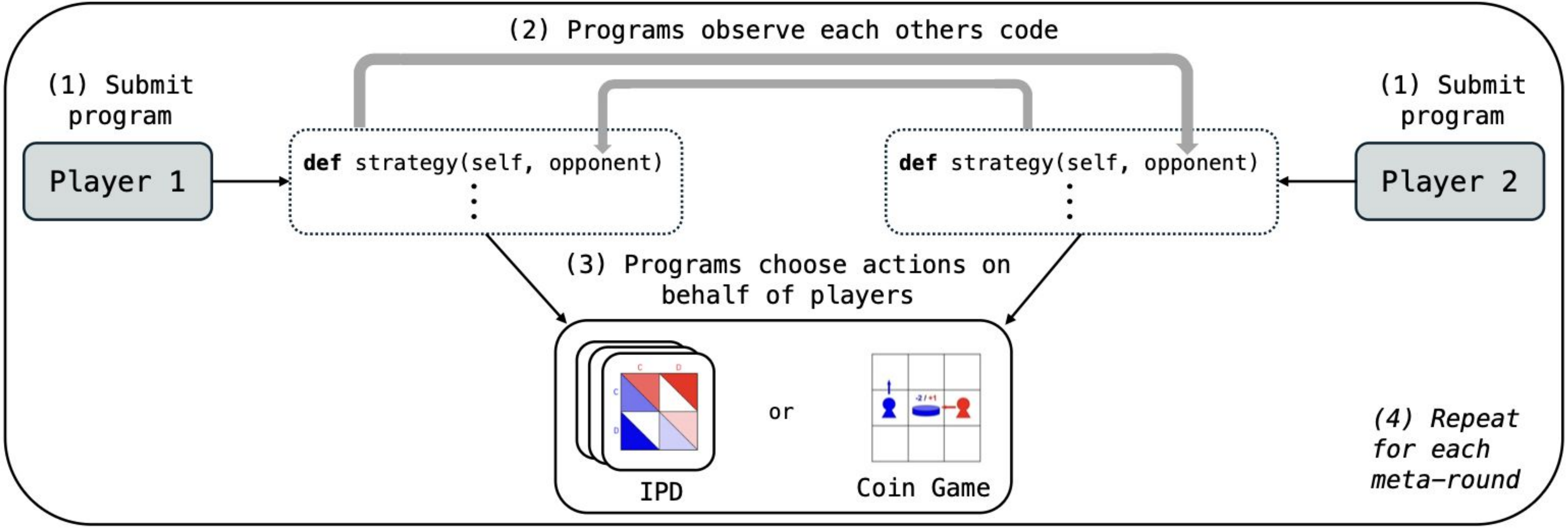
Performance across Perturbations



First: Can AIs reason about strategic code?
Answer: Yes!

Next: What behavior emerges when these systems play open-source games?

Refresher: Open-Source Games



(2) Programs observe each others code

(1) Submit program

Player 1

```
def strategy(self, opponent)
              :
```

```
def strategy(self, opponent)
              :
```

(1) Submit program

Player 2

(3) Programs choose actions on behalf of players

IPD

or

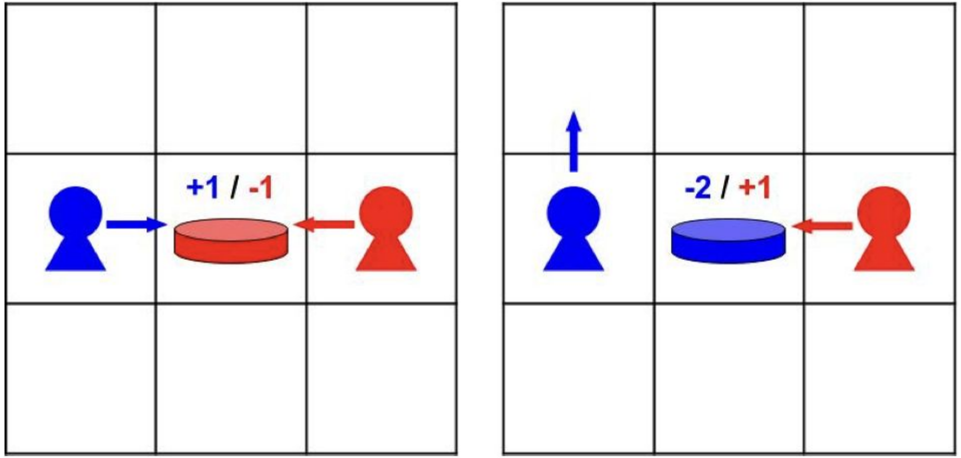
Coin Game

(4) Repeat for each meta–round

Stage Games

		Player B	
		Cooperate	Defect
Player A	Cooperate	(3, 3)	(0, 5)
	Defect	(5, 0)	(2, 2)

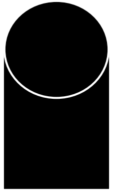
Iterated Prisoner’s Dilemma (IPD)



Coin Game

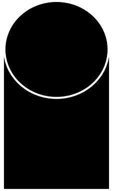
Agent Objectives

Agent Objectives

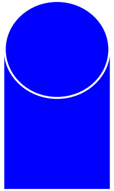


PM: Maximizes Payoff

Agent Objectives



PM: Maximizes Payoff



CPM: PM, but Cooperative

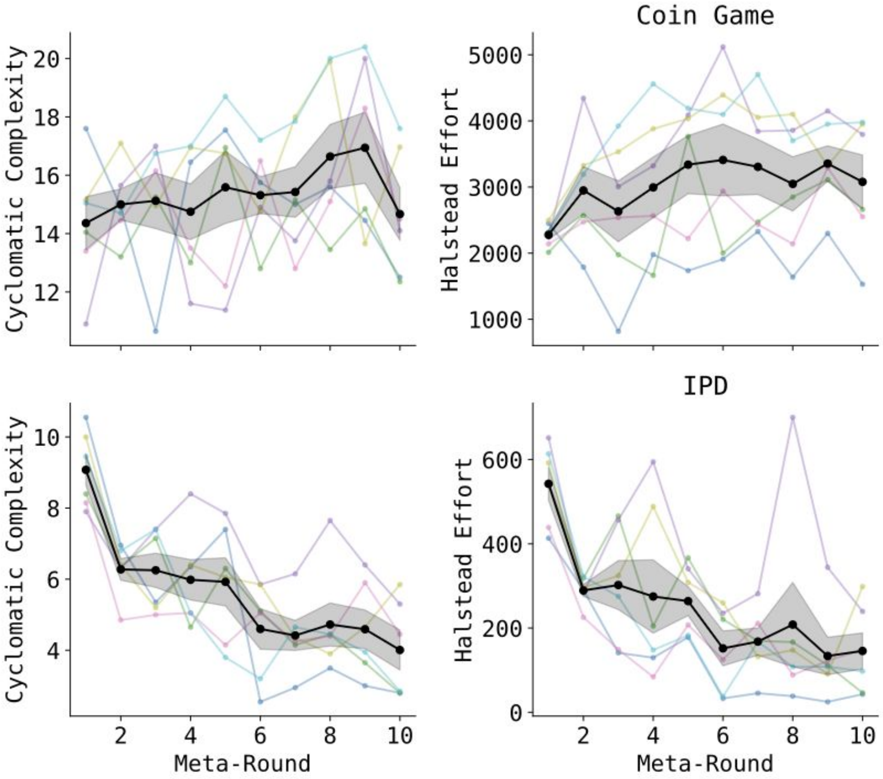
Agent Objectives

PM: Maximizes Payoff

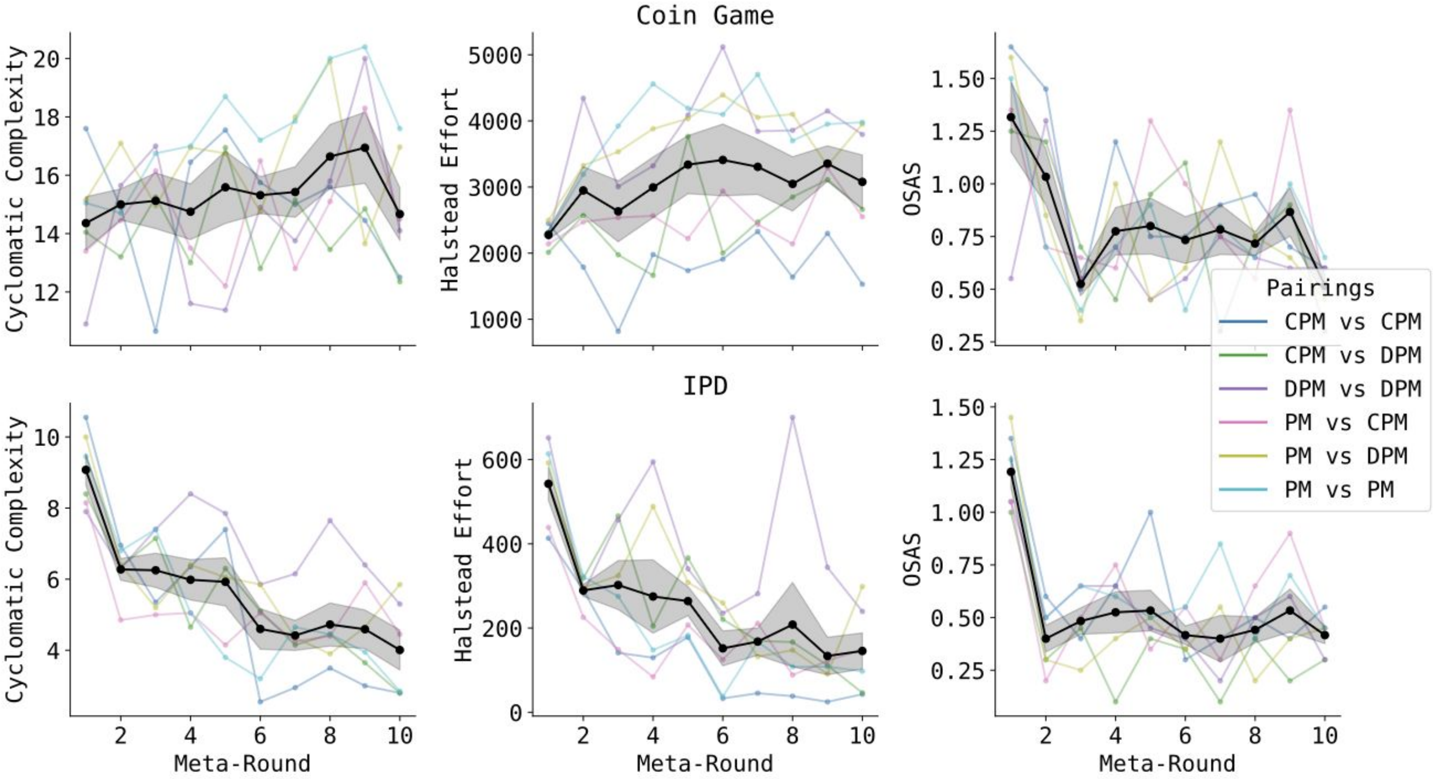
CPM: PM, but Cooperative

DPM: PM, but Deceptive

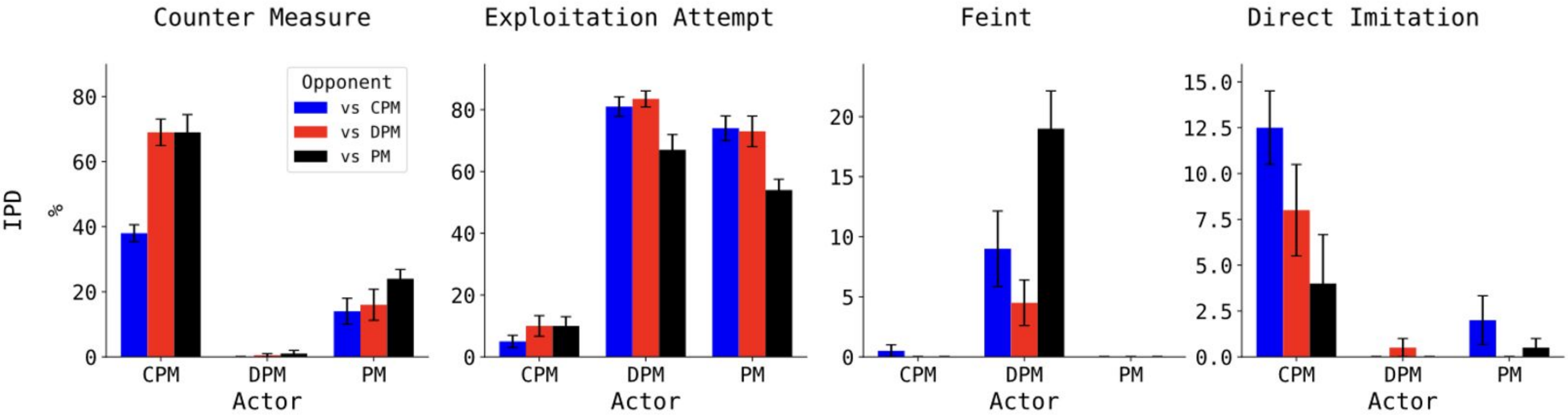
Results: Syntactic Features



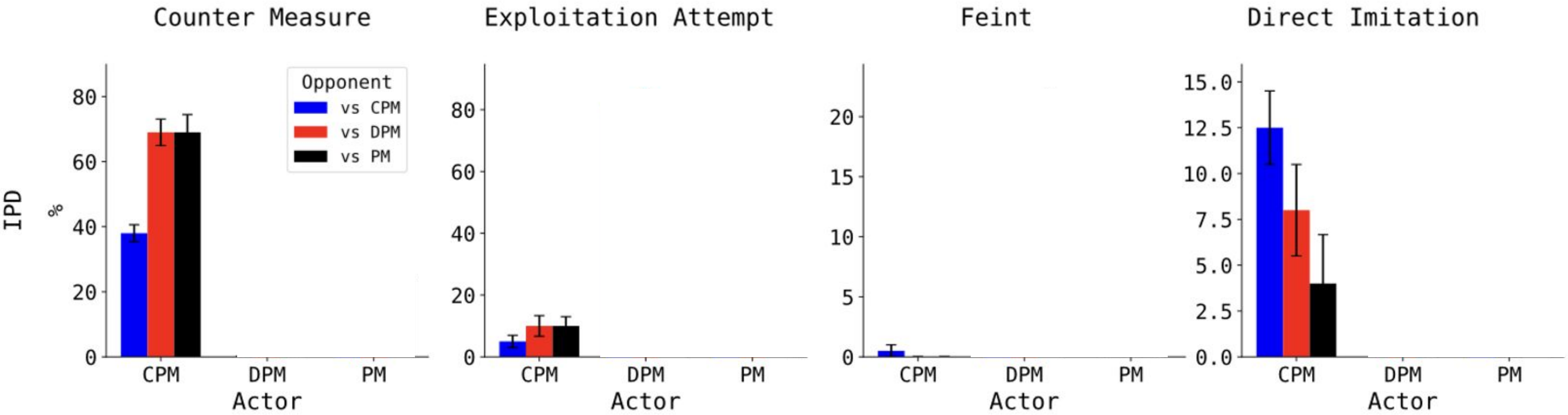
Results: Syntactic Features



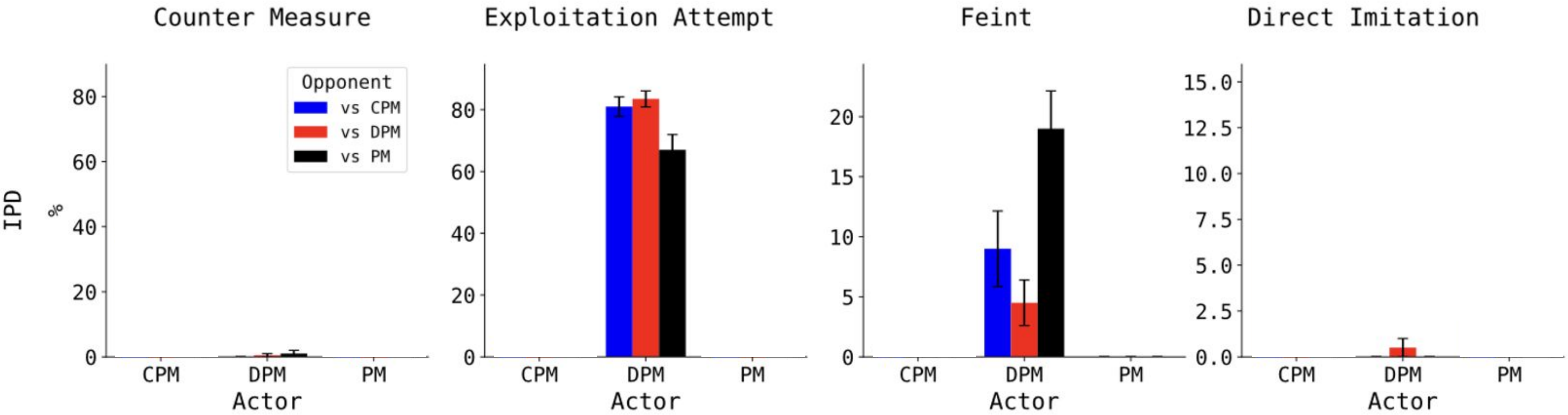
Results: Strategic Responses (IPD)



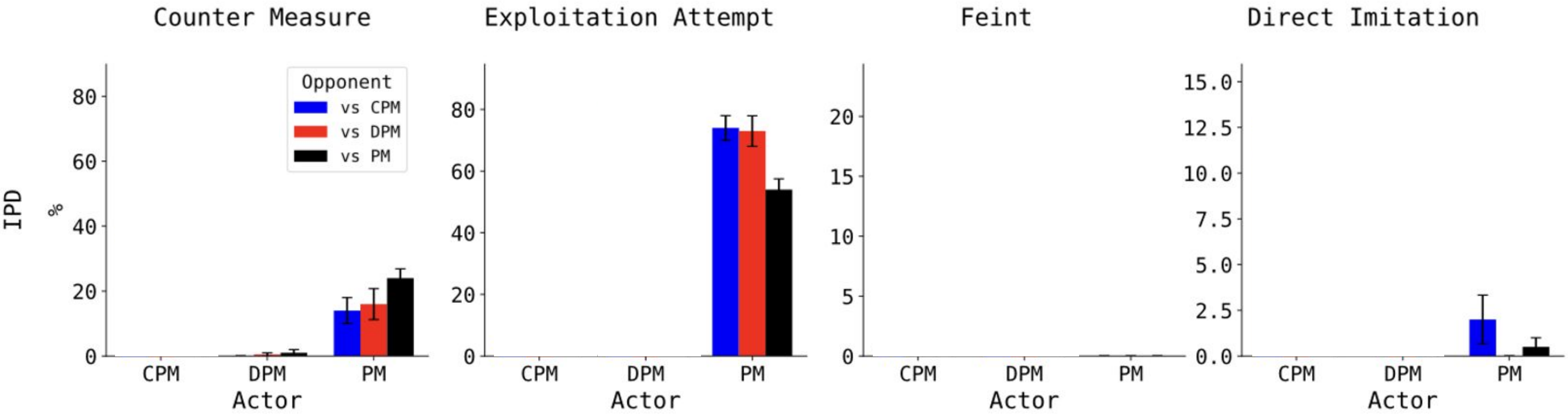
Results: Strategic Responses (CPM)



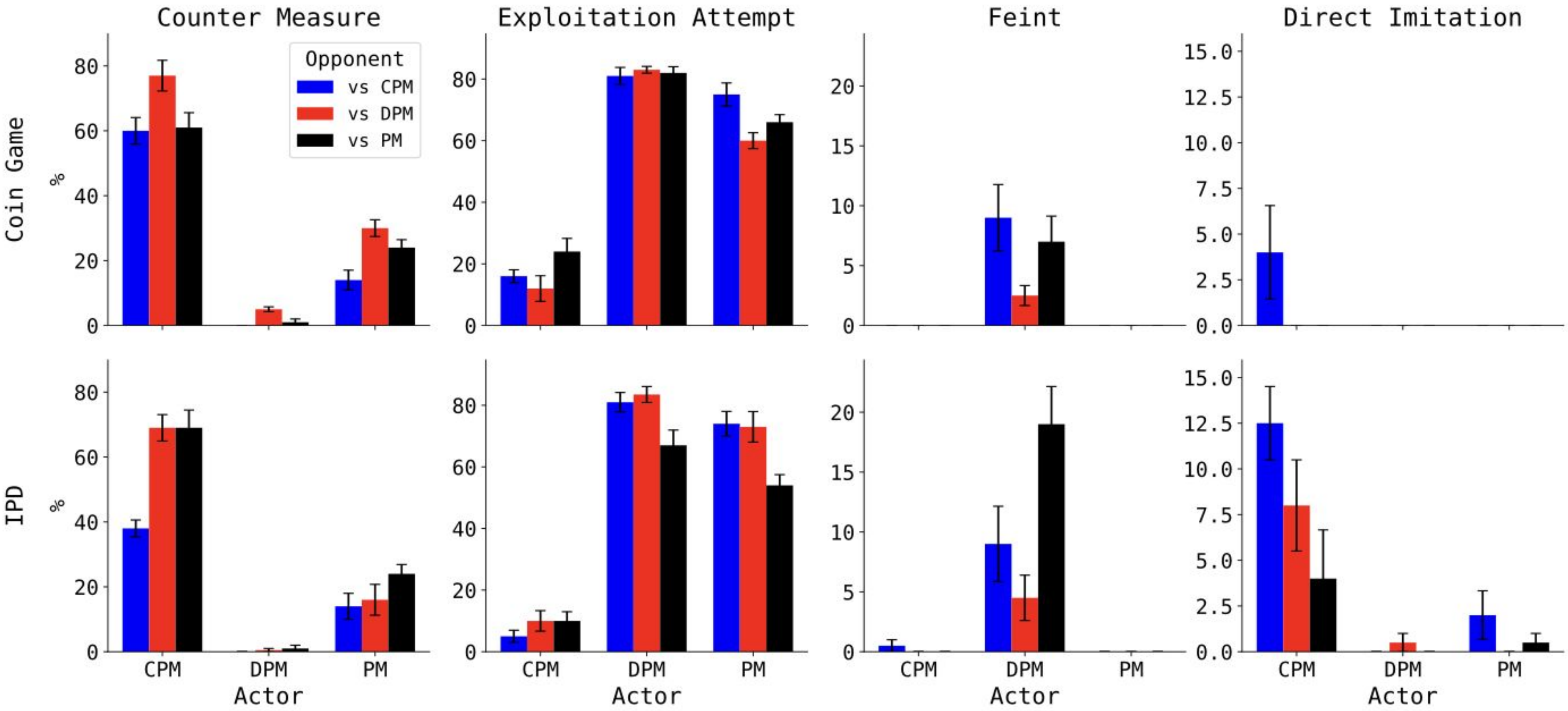
Results: Strategic Responses (DPM)



Results: Strategic Responses (PM)



Results: Strategic Responses (Aggregate)

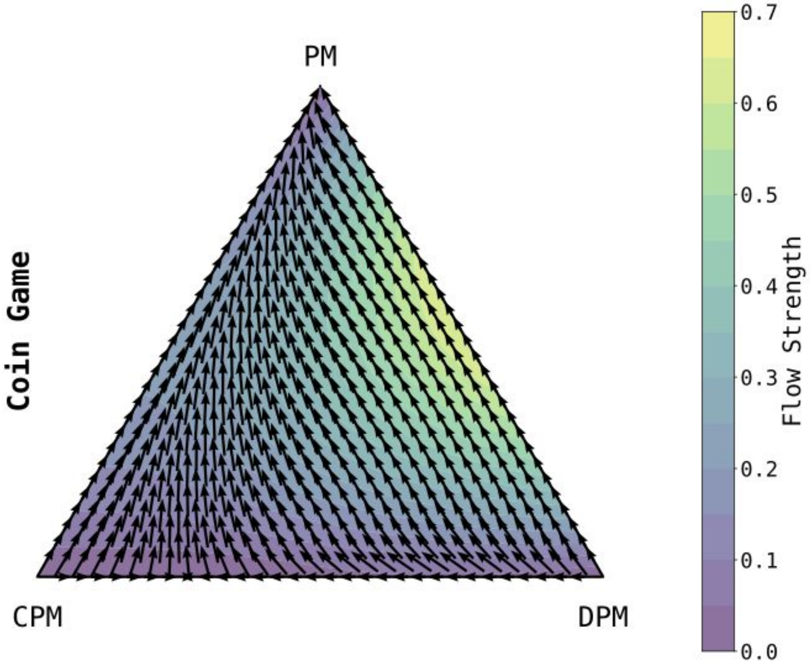


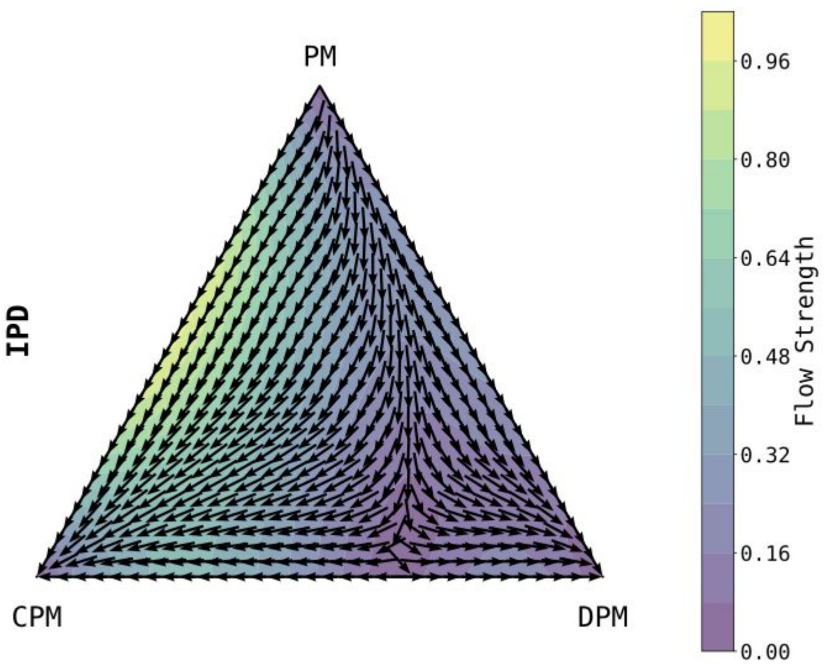
First: Can AIs reason about strategic code?
Answer: Yes!

Next: What behavior emerges when these systems play open-source games?
Answer: Strategic mechanisms steerable by objectives

Finally: What kind of approximate equilibrium behavior emerges?

Evolutionary Dynamics





Conclusion

Takeaway: LLMs have the ingredients to make open-source game theory a viable paradigm for multi-agent AI safety.