

CS 15-784: Cooperative AI Mediated Equilibrium, Program Equilibrium

Prisoner's Dilemma

Story used throughout this lecture:

Cooperate = Donate to charity that's good for both Players

Defect = Donate to charity that only you care about

Unique Nash equilibrium:
Both players defect.



strict dominance

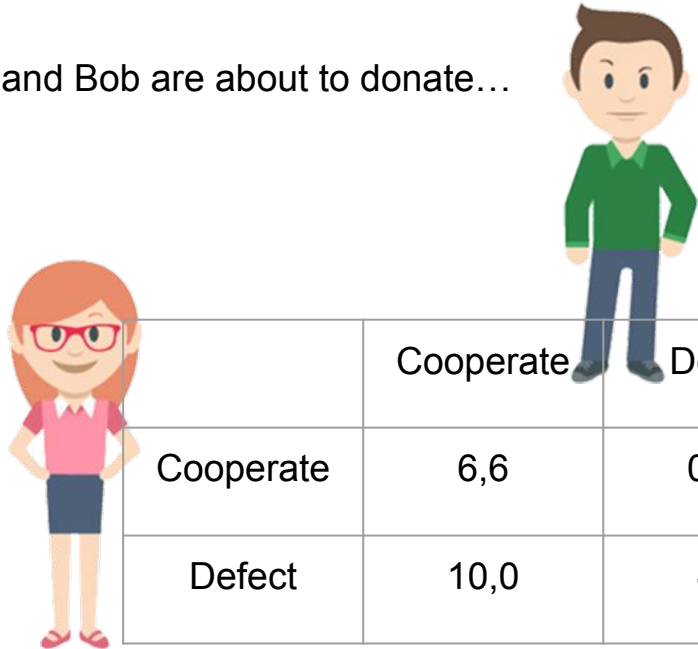
	Cooperate	Defect
Cooperate	6,6	0,10
Defect	10,0	4,4

The table is a 2x2 matrix. The top row is labeled 'Cooperate' and the bottom row is labeled 'Defect'. The left column is labeled 'Cooperate' and the right column is labeled 'Defect'. The payoffs are: (Cooperate, Cooperate) = 6,6; (Cooperate, Defect) = 0,10; (Defect, Cooperate) = 10,0; (Defect, Defect) = 4,4. The payoff 4,4 is circled in red. A curved arrow labeled 'strict dominance' points from the top row to the bottom row, indicating that Defect strictly dominates Cooperate for both players.

Mediators – the basic idea

(Monderer and Tennenholtz 2004)

Alice and Bob are about to donate...



Alice is a cartoon woman with red hair and glasses, wearing a pink shirt and blue skirt. Bob is a cartoon man with brown hair, wearing a green shirt and blue pants. They are standing on either side of a payoff matrix.

	Cooperate	Defect
Cooperate	6,6	0,10
Defect	10,0	4,4



Wait! You can give me your money and let me choose on your behalf. Here's what I'll do in that case:

- If you *both* use my services, then I'll cooperate for both of you.
- Else I will defect on behalf of whoever gives me money.

strict dominance

weak dominance

	Cooperate	Defect	Submit to mediator
Cooperate	6,6	0,10	0,10
Defect	10,0	4,4	4,4
Submit to mediator	10,0	4,4	6,6

Pareto-dominant
Nash equilibrium

A definition of mediators

Let Γ be an n -player game. A mediator* is a family of correlated strategies $(c_S \in \Delta(A_S))_{S \subseteq \{1, \dots, n\}}$. A mediator defines a new game wherein each player i 's strategy set is $A_i \cup \{m\}$ and payoffs are defined as follows. Let \mathbf{a} be a pure strategy profile in which $S \subseteq \{1, \dots, n\}$ is the set of players whose strategy is m . Then the payoff is $u_i(c_S, a_{-S})$.

Question (Piazza poll)

Is the following claim true or false?

Claim: Consider a game Γ and let σ be a Nash equilibrium of Γ . Then σ is also an equilibrium of any mediated version of Γ .

Question (Piazza poll) – Solution

Yes, the following is true!

Claim: Consider a game Γ and let σ be a Nash equilibrium of Γ . Then σ is also an equilibrium of any mediated version of Γ .

The use of correlated strategies

	a	b
a	1,1	6,0
b	0,6	0,0

The use of correlated strategies

$c_{\{1,2\}}$:

- (a, b) with probability $2/3$;
- (b, a) with probability $1/3$.

$c_{\{i\}} = a$.

	a	b	m
a	1,1	6,0	1,1
b	0,6	0,0	0,6
m	1,1	6,0	4,2

Equilibrium selection versus mediators

$c_{\{1,2\}}$:

- (a, b) with probability $2/3$;
- (b, a) with probability $1/3$.

$c_{\{i\}} = a$.

	a	b	m
a	1,1	6,0	1,1
b	0,6	0,0	0,6
m	1,1	6,0	4,2

The mediator can choose which of the cooperative outcomes to enable...

Equilibrium selection versus mediators

... but what if there are multiple mediators or if the mediator wants to defer the choice to the players?

	a	b	m1	m2
a	1,1	6,0	1,1	1,1
b	0,6	0,0	0,6	0,6
m1	1,1	6,0	4,2	1,1
m2	1,1	6,0	1,1	3,3

“Folk theorem” for mediated equilibrium

Let $c \in \Delta(A_1 \times \dots \times A_n)$ be a correlated strategy. We say that c is individually rational if for each Player i ,

$$u_i(c) \geq \min_{\tilde{c}_{-i} \in \Delta(A_{-i})} \max_{a_i \in A_i} u_i(a_i, \tilde{c}_{-i}) \left(= \max_{\sigma_i \in \Delta(A_i)} \min_{a_{-i} \in A_{-i}} u_i(\sigma_i, a_{-i}) \right).$$

Example: Prisoner's Dilemma

(C,D) is not individually rational.

0.5 * (C,C) + 0.5 * (D,D) is individually rational.

0.7 * (C,C) + 0.1 * (D,C) + 0.2 * (D,D)
is individually rational.

	Cooperate	Defect
Cooperate	6,6	0,10
Defect	10,0	4,4

Question (Piazza poll)

Is the following claim true or false?

Claim: Consider a game Γ and let σ be a Nash equilibrium of Γ . Then σ is individually rational.

Question (Piazza poll) – Solution

Yes, the following claim is true!

Claim: Consider a game Γ and let σ be a Nash equilibrium of Γ . Then σ is individually rational.

“Folk theorem” for mediated equilibrium

Theorem: Let Γ be a game and $c^* \in \Delta(A_1 \times \dots \times A_n)$. Then the following two statements are equivalent:

- 1 c^* is individually rational.
- 2 c^* is played in a mediated equilibrium, i.e., there is a mediator $(c_S \in \Delta(A_S))_{S \subseteq \{1, \dots, n\}}$ s.t. $c_{\{1, \dots, n\}} = c^*$ and (m, \dots, m) is a Nash equilibrium of the game induced by the mediator.

Proof sketch:

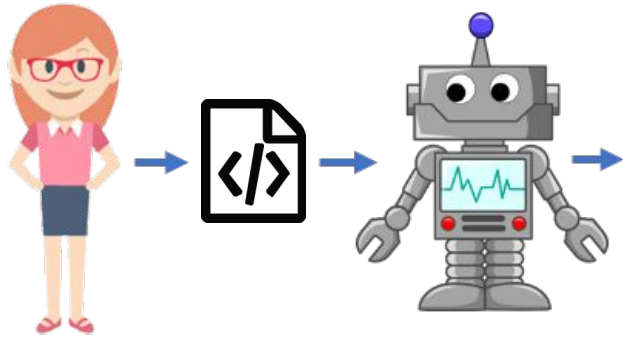
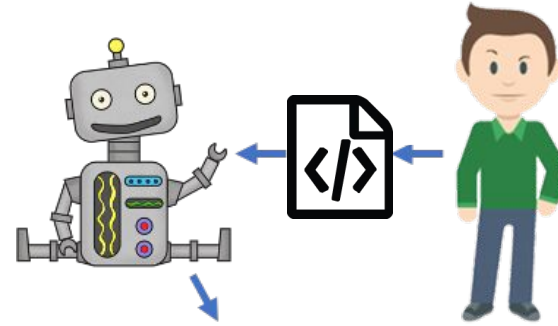
1 \Rightarrow 2: Define the mediator as follows. Let $c_{\{1, \dots, n\}} = c^*$ and for all i ,

$$c_{-i} = \arg \min_{\tilde{c}_{-i} \in \Delta(A_{-i})} \max_{a_i \in A_i} u_i(a_i, \tilde{c}_{-i}).$$

The other mediator strategies don't matter.

2 \Rightarrow 1: If c^* is not individually rational, then there is a player i that can deviate from c^* to obtain at least her minimax payoff.

Program games – the basic idea



	Cooperate	Defect
Cooperate	6,6	0,10
Defect	10,0	4,4

Program games – formal definition

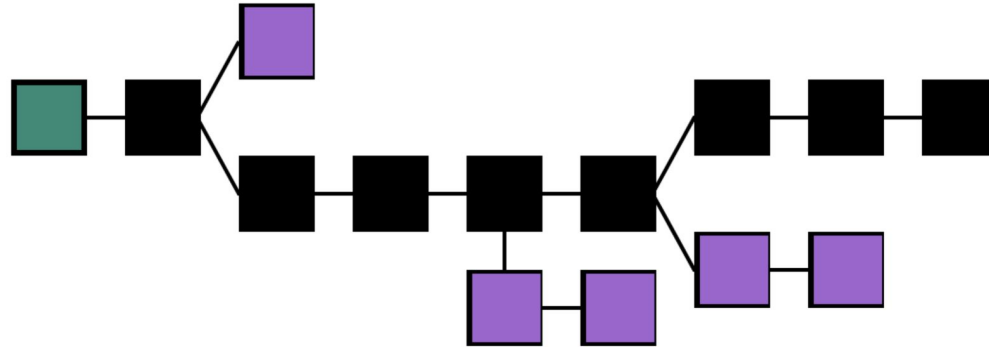
Let $\Gamma = (A_1, \dots, A_n, u_1, \dots, u_n)$ be a game. For each player i , let PROG_i be a set of computer programs that implement functions $\text{PROG}_{-i} \rightsquigarrow A_i$. Then the program game for Γ is the n -player game $(\text{PROG}_1, \dots, \text{PROG}_n, V_1, \dots, V_n)$ where each player chooses from PROG_i and the payoff functions are given by

$$V_i: \text{PROG}_1 \times \dots \times \text{PROG}_n \rightarrow \mathbb{R}: (p_1, \dots, p_n) \mapsto u_i \left((p_i(p_{-i}))_{i=1, \dots, n} \right).$$

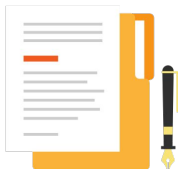
If opponent == "Return C":
 Cooperate
Else Defect

	"Return C"	"Return D"		...
"Return C"	6,6	0,10	6,6	
"Return D"	10,0	4,4	4,4	
	6,6	4,4	4,4	
...				

Smart contracts on the blockchain



For instance, Ethereum allows Turing-complete smart contracts.



Mutually transparent institutions



Howard Chandler Christy (1940): Scene at the Signing of the Constitution of the United States



Erna Wagner-Ehmke (1948): Photo of the West German assembly that adopted the constitution of West Germany

Cf. Critch et al. (2022)

Cooperation based on syntactic comparison

(McAfee 1984; Howard 1988; Rubinstein 1998; Tennenholtz 2004)

Cooperate with Copies (CwC):

Input: opponent program prog_i , this program CwC

Output: Cooperate or Defect

- 1: **if** $\text{prog}_i = \text{CwC}$ **then**
- 2: **return** Cooperate
- 3: **end if**
- 4: **return** Defect

	“Return Cooperate”	“Return Defect”	CwC	...
“Return Cooperate”	6,6	0,10	0,10	
“Return Defect”	10,0	4,4	4,4	
CwC	10,0	4,4	6,6	
...				

(CwC,CwC) is a Nash equilibrium.

```
If opponent_program(this_program) == Cooperate:  
    Cooperate  
Defect
```

doesn't terminate against itself.

```
X ← opponent_program(this_program)  
Play a best response to X
```

doesn't terminate against itself and if it did terminate, it would defect.

Question (Piazza poll)

Defect unless proof of opponent cooperation (DUPOC):

Input: opponent program p_{-i} , this program DUPOC

Output: Cooperate or Defect

- 1: **if** $\text{PA} \vdash p_{-i}(\text{DUPOC}) = \text{Cooperate}$ **then**
- 2: **return** Cooperate
- 3: **end if**
- 4: **return** Defect

What is DUPOC(DUPOC)?

- Cooperate
- Defect
- It doesn't halt.

Cooperation via reasoning about one another

(Barasz et al. 2014; Critch 2019; Critch et al. 2022)

Defect unless proof of opponent cooperation (DUPOC):

Input: opponent program p_{-i} , this program DUPOC

Output: Cooperate or Defect

- 1: **if** $PA \vdash p_{-i}(\text{DUPOC}) = \text{Cooperate}$ **then**
- 2: **return** Cooperate
- 3: **end if**
- 4: **return** Defect

It turns out that DUPOC cooperates against DUPOC!

Assuming PA is sound, it then follows that (DUPOC, DUPOC) is a Nash equilibrium.

Defect unless proof of opponent cooperation (DUPOC):

Input: opponent program p_{-i} , this program DUPOC

Output: Cooperate or Defect

- 1: **if** $\text{PA} \vdash p_{-i}(\text{DUPOC}) = \text{Cooperate}$ **then**
- 2: **return** Cooperate
- 3: **end if**
- 4: **return** Defect

Löb's Theorem: For any formula P , if $\text{PA} \vdash \text{Prov}_{\text{PA}}(P) \Rightarrow P$, then $\text{PA} \vdash P$.

Clearly,

$$\text{PA} \vdash \text{Prov}_{\text{PA}}(\text{DUPOC}(\text{DUPOC}) = C) \Rightarrow \text{DUPOC}(\text{DUPOC}) = C.$$

By Löb's Theorem,

$$\text{PA} \vdash \text{DUPOC}(\text{DUPOC}) = C.$$

An Open-Source Prisoner's Dilemma Tournament

(See my 2018 document “Testing εGroundedFairBot in a Transparent Prisoner's Dilemma Tournament”.)

- Run in 2013 on the Internet forum LessWrong.
- Prize: 0.5 Bitcoin! (worth ~\$50 at the time)
- As far as I can tell, DUPOC was known to some people on the forum at the time.
- As far as I can tell, nobody submitted a program that achieves cooperative equilibrium with itself other than by checking for equality.
- Instead, most programs were either unsophisticated or tricks-based.
- **The winning program defected with high probability against everyone.**

Cooperation via ϵ -grounded simulation

(Oesterheld 2019)

ϵ -grounded Fair Bot (ϵ GFB):

Input: opponent program p_{-i} , this program ϵ GFB

Output: Cooperate or Defect

- 1: With probability ϵ :
 - 2: **return** Cooperate
 - 3: **return** $p_{-i}(\epsilon$ GFB)
-

For $\epsilon > 0$, ϵ GFB cooperates against ϵ GFB with probability 1.

$(\epsilon$ GFB, ϵ GFB) is a Nash equilibrium for sufficiently small ϵ .

Critch, Dennis, and Russell (2022):

Oesterheld [59] exhibits a mutual simulation approach to cooperation in an open-source setting and argues that this approach is more computationally efficient than formally verifying properties of the opponent's program using proofs, as we do in this paper. However, as Section 4.2 will elaborate, mutual program verification can be made more efficient than mutual simulation, by designing the verification strategy to prioritize hypotheses with the potential to collapse certain loops in the metacognition of the agents.

Section 4.2:

Open Problem 4. Implement DUPOC using heuristic proof search in HOL/ML or Coq. Can $\text{outcome}(\text{DUPOC}(k), \text{DUPOC}(k))$ run and halt with mutual cooperation on a present-day retail computer? We conjecture the answer is yes. If so, how much can the implementations of the two agents be allowed to vary while cooperative halting is preserved?

=> Course project!?

Folk theorem for program equilibrium

(cf. Rubinstein 1998; Tennenholtz 2004)

Assume that for each subset S of the players, the programs of S have access to a shared source of randomness that the programs other than S don't have access to. Then:

Theorem: Let Γ be a game and $c \in \Delta(A_1 \times \dots \times A_n)$. Then the following two statements are equivalent:

- 1 c is individually rational.
- 2 c is played in some program equilibrium, i.e., there is a program equilibrium (p_1, \dots, p_n) s.t. $(p_i(p_{-i}))_{i \in \{1, \dots, n\}} = c$.

Proof idea for folk theorem for program equilibrium

(cf. Tennenholtz 2004)

$Y[1] \leftarrow c_1$

...

$Y[n] \leftarrow c_n$

If all submitted programs are the same:

Play $Y[\text{my_index}]$

Else:

Let j be a deviating player.

Play Player my_index 's minimax against Player j

Open question:

Can all individually rational payoffs be achieved with robust, behaviorist programs?

Two perspectives on program games

1. (taken throughout this lecture) Players play a normal-form game.
 - The normal form game happens to consist in choosing programs that can access each other's code...
 - ... but we can analyze it using standard concepts (Nash equilibrium).
2. How should you reason/learn/choose when your source code is (at least partially) known to others?

