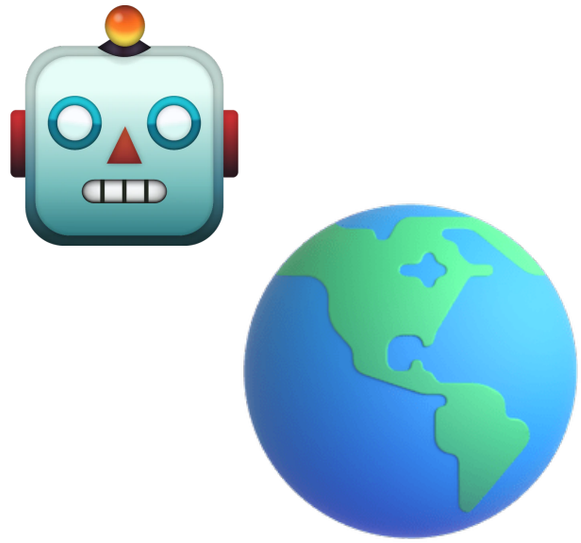
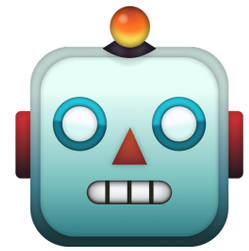
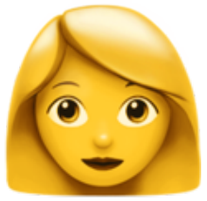
Guarantees for Self-Play via Polymatrix Decomposability

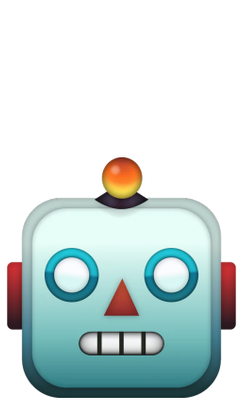
Revan MacQueen

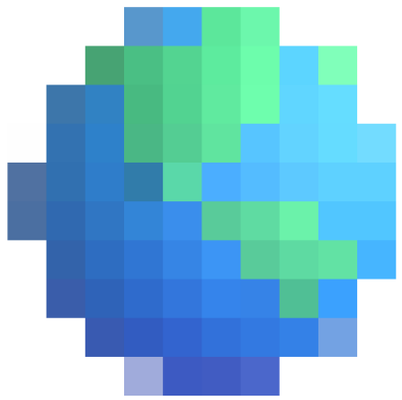
Thesis Presentation. August 28, 2023

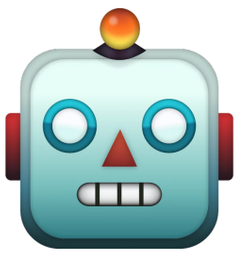


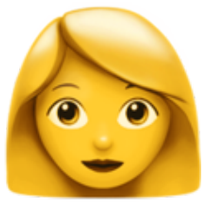


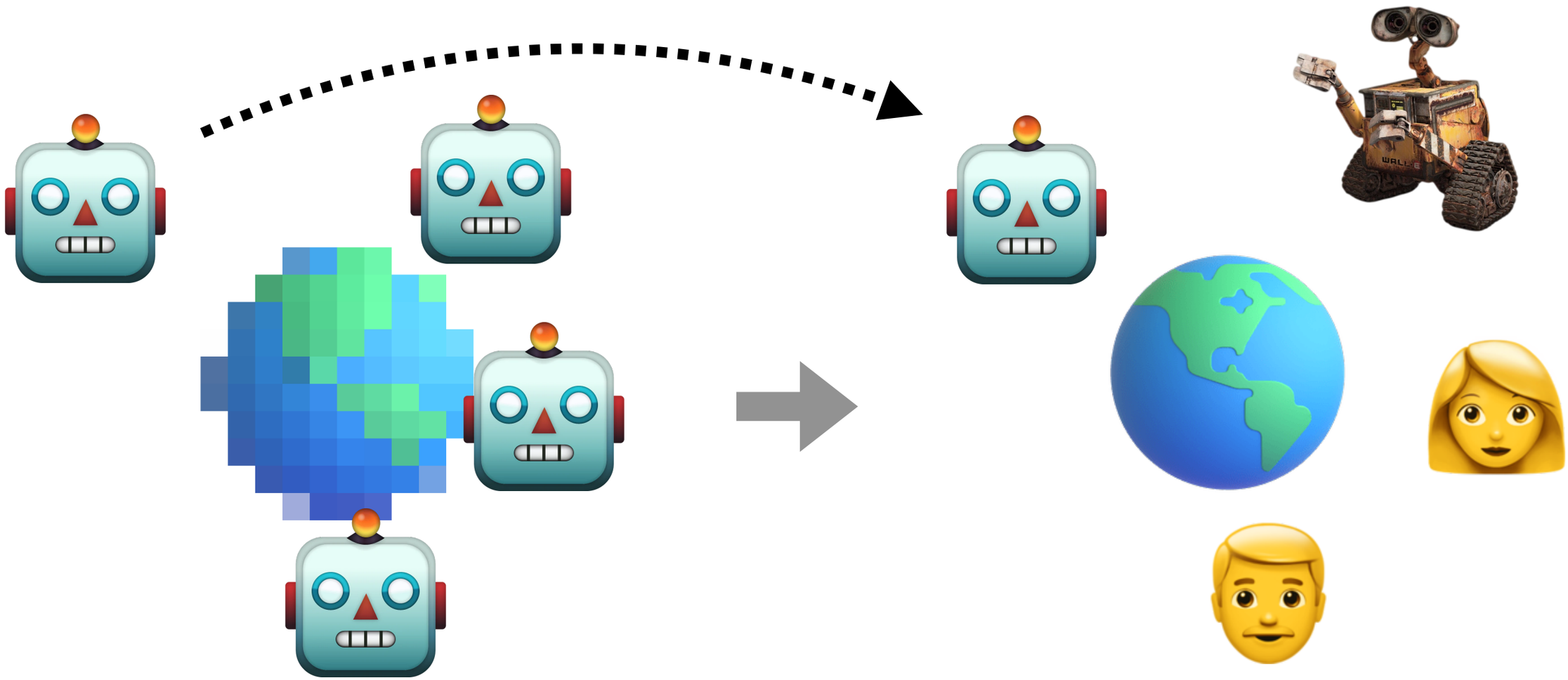






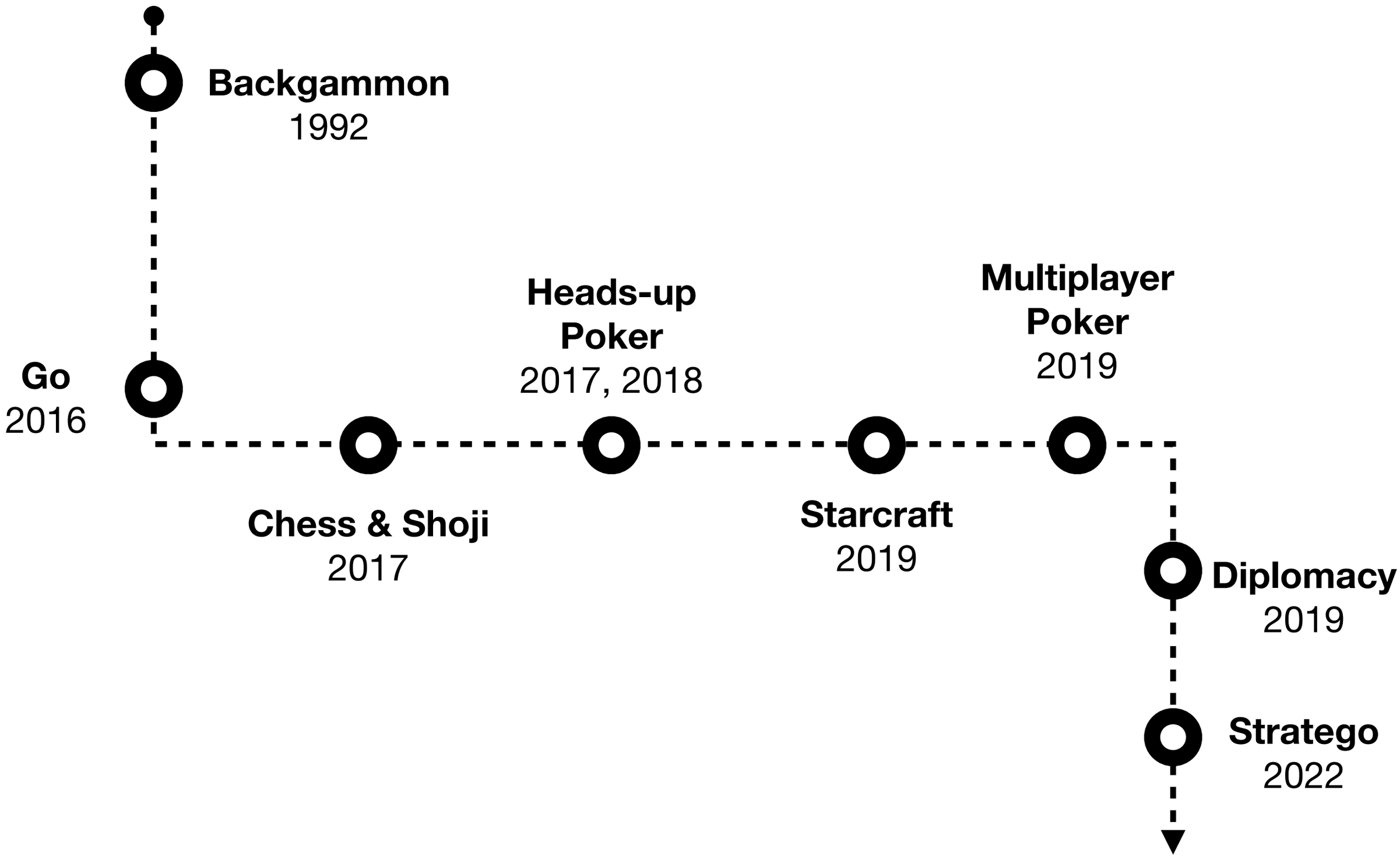




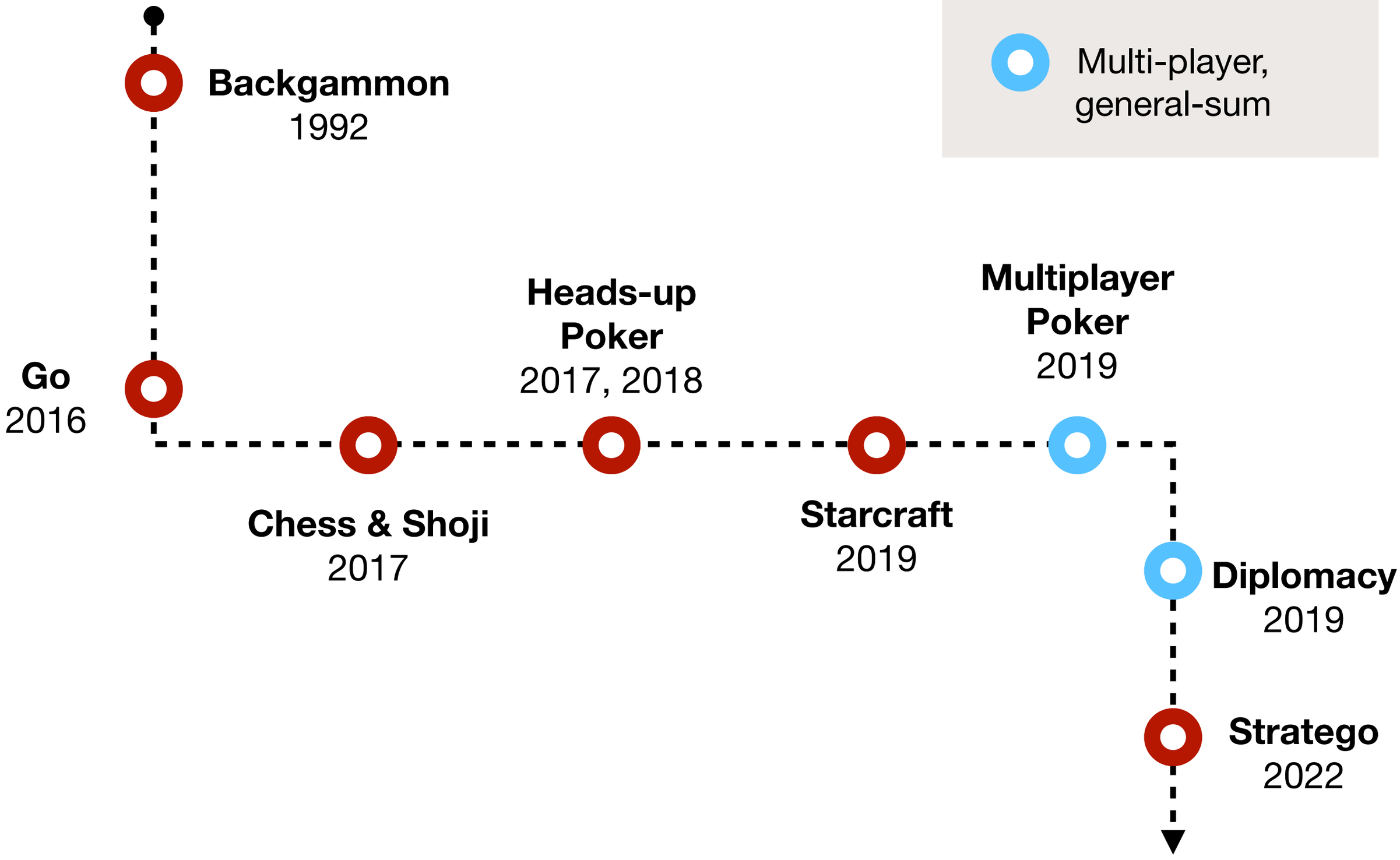


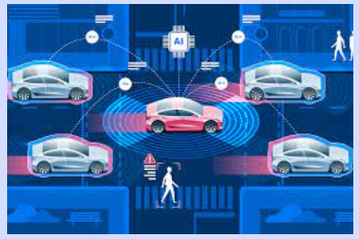
Self-Play

Success of Self-Play



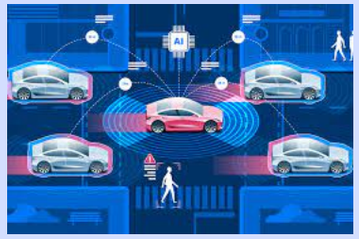
Success of Self-Play





multi-player,
general-sum

Two-player,
constant-sum

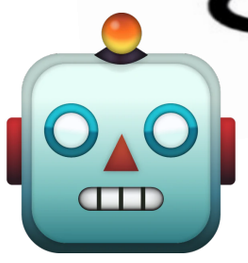


?

Two-player,
constant-sum

multi-player,
general-sum

What are properties of games where self-play produces **good** strategies?

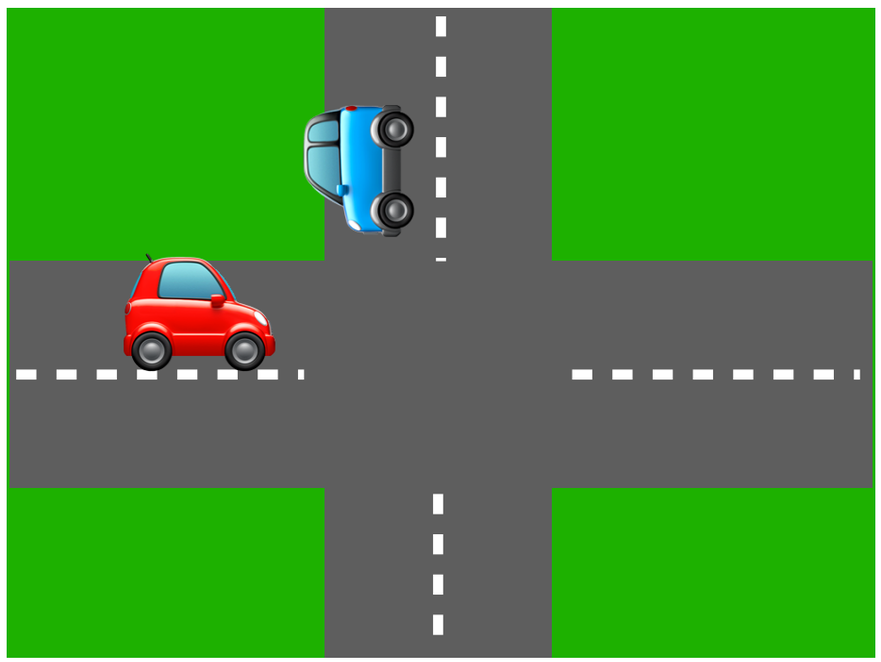


Background

Game Theory Basics

Normal-Form Games

Agents	Pure strategies	Mixed Strategies
N	P_i	S_i
🚘 🚙	$\{\text{go}, \text{stop}\}$	$s_i \in \Delta(\{\text{go}, \text{stop}\})$



Utility Functions

Pure strategy **profiles**

P

$(\text{go}, \text{stop}) \in P$

Mixed strategy **profiles**

S

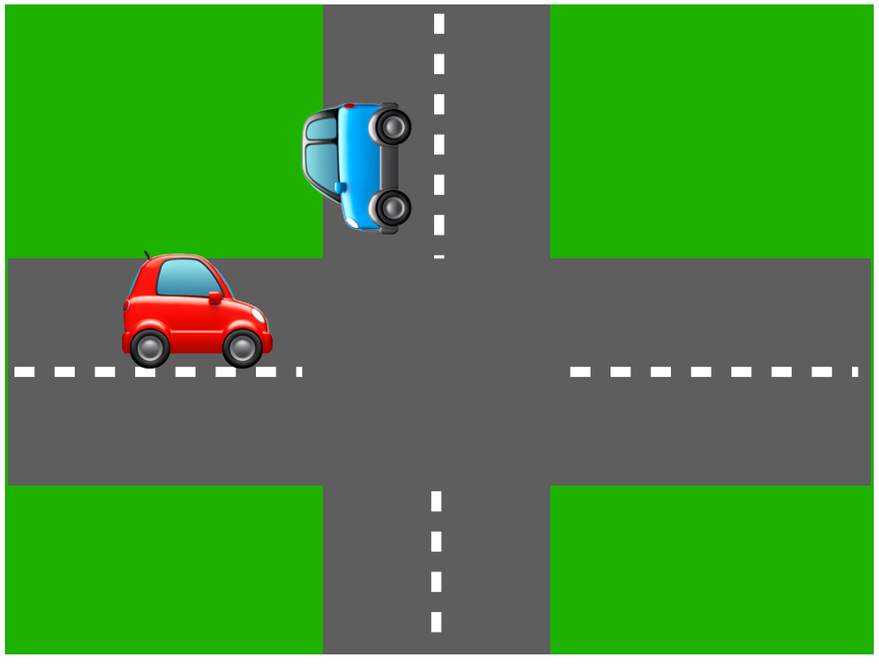
$s \in S$

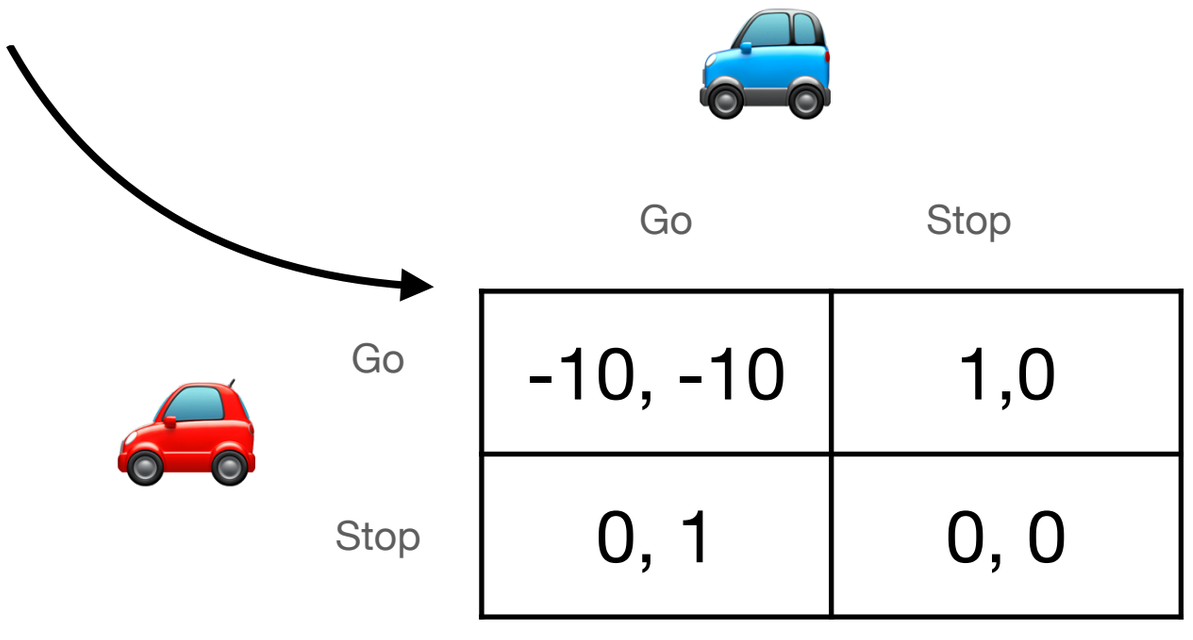
Utility functions

$u_i : P \to \mathbb{R}$

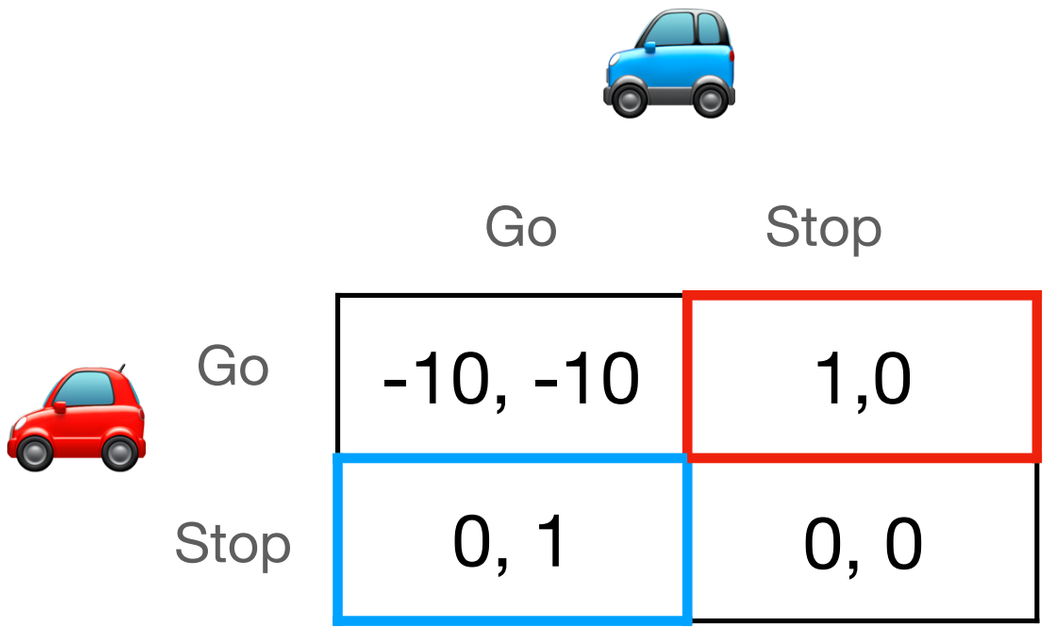
$u_i(s) \doteq \mathbb{E}_{\rho \sim s}[u_i(\rho)]$

No optimal strategy!



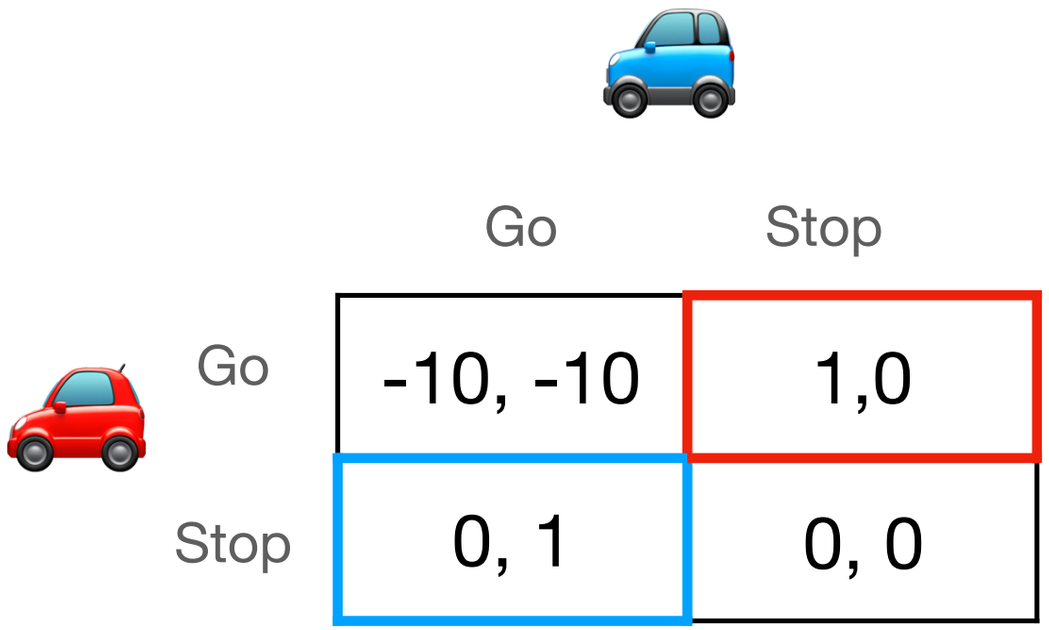


Nash Equilibrium



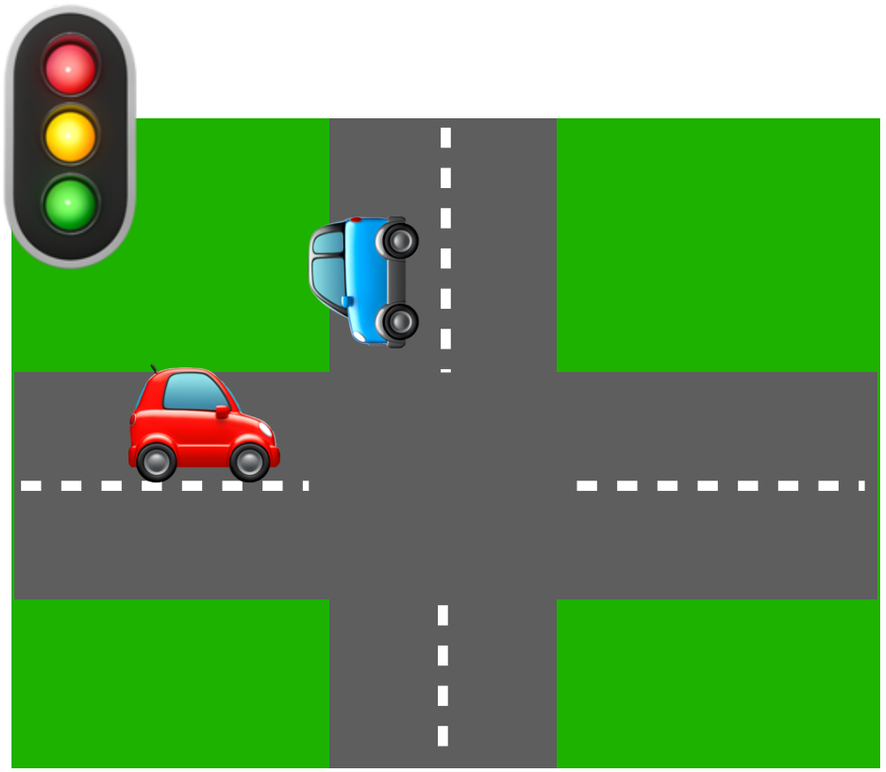
Definition. A Nash equilibrium is a **strategy profile** such that no players wish to deviate to another strategy.

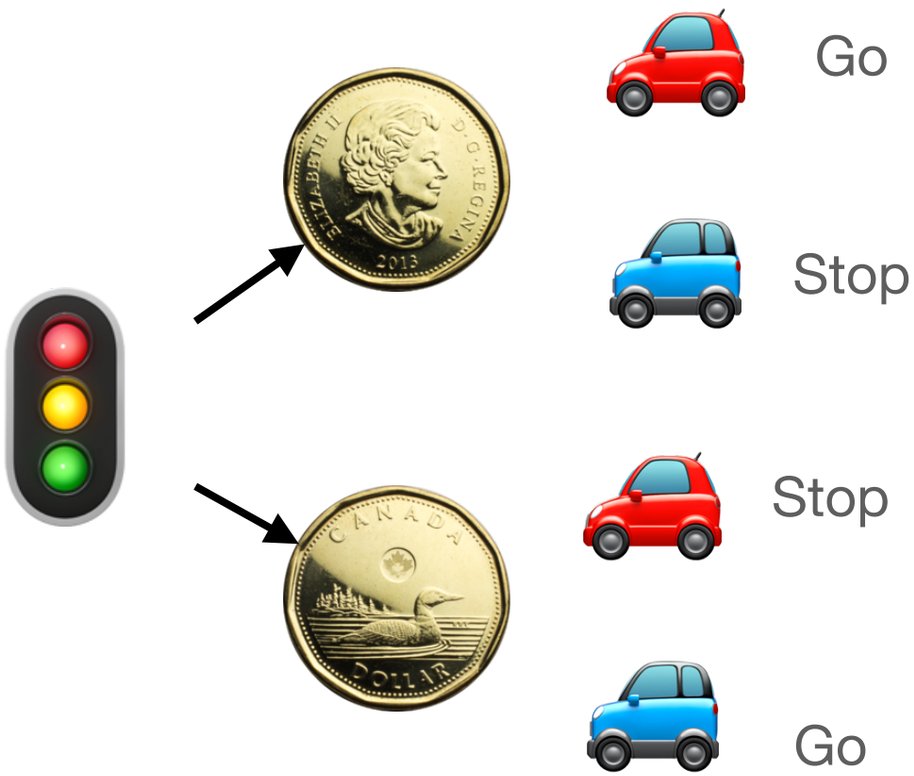
Approximate Nash Equilibrium



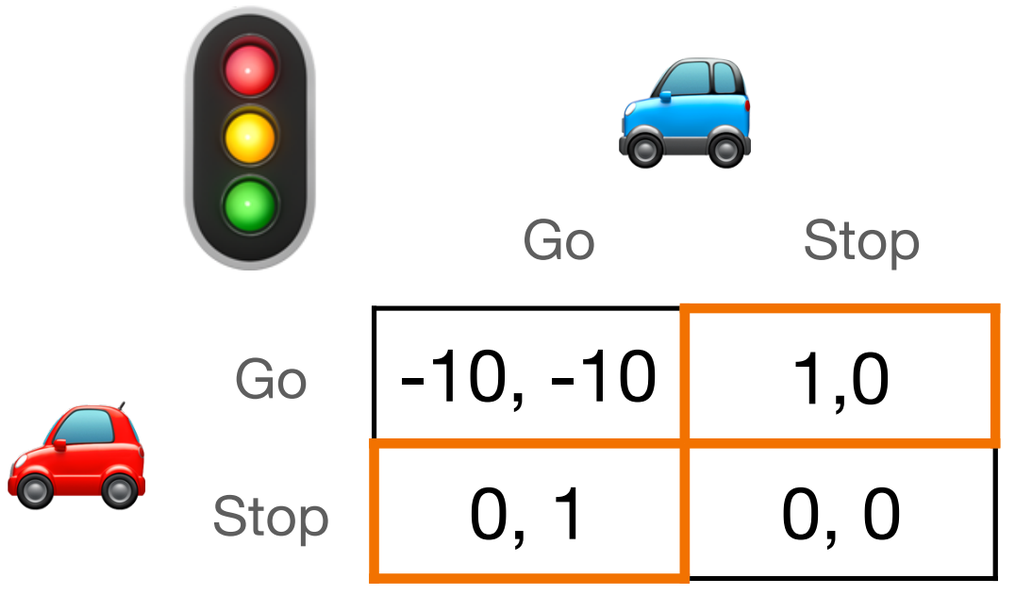
Definition. A **strategy profile** is an ϵ-Nash equilibrium if no player can gain more than ϵ utility by deviating.

Coarse Correlated Equilibrium (CCE)





Coarse Correlated Equilibrium (CCE)

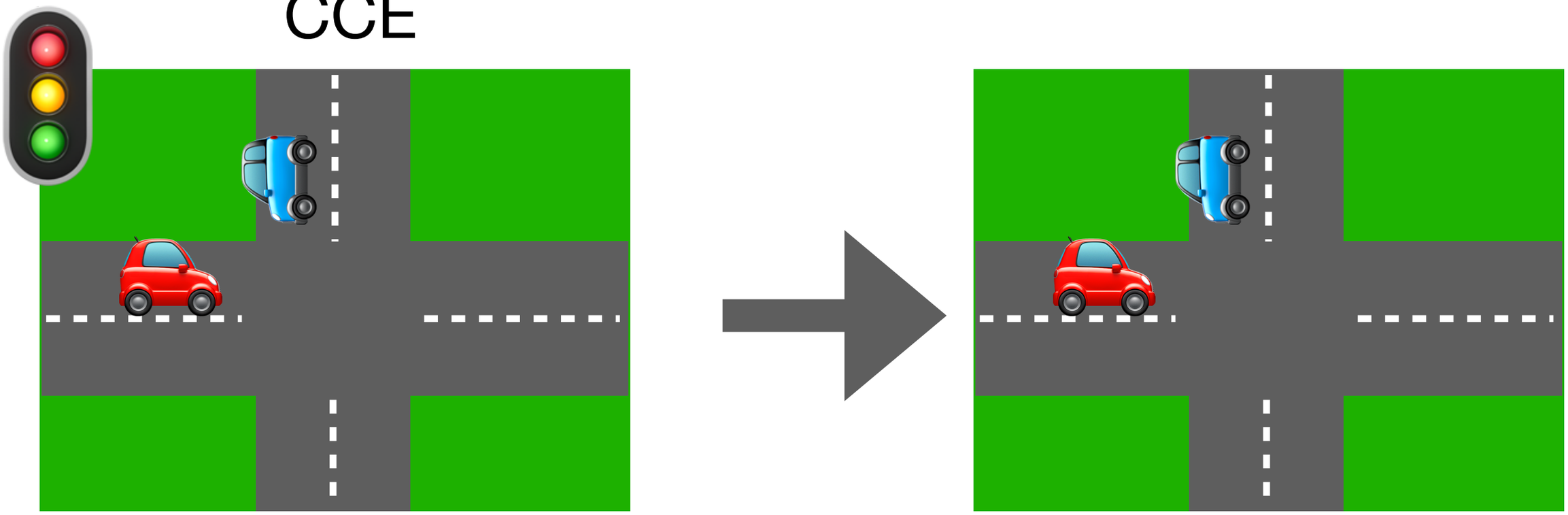


Each player gets expected utility of 0.5

No player wishes to **deviate** from their recommendations

What If Players Lose the Ability to Correlate?

CCE

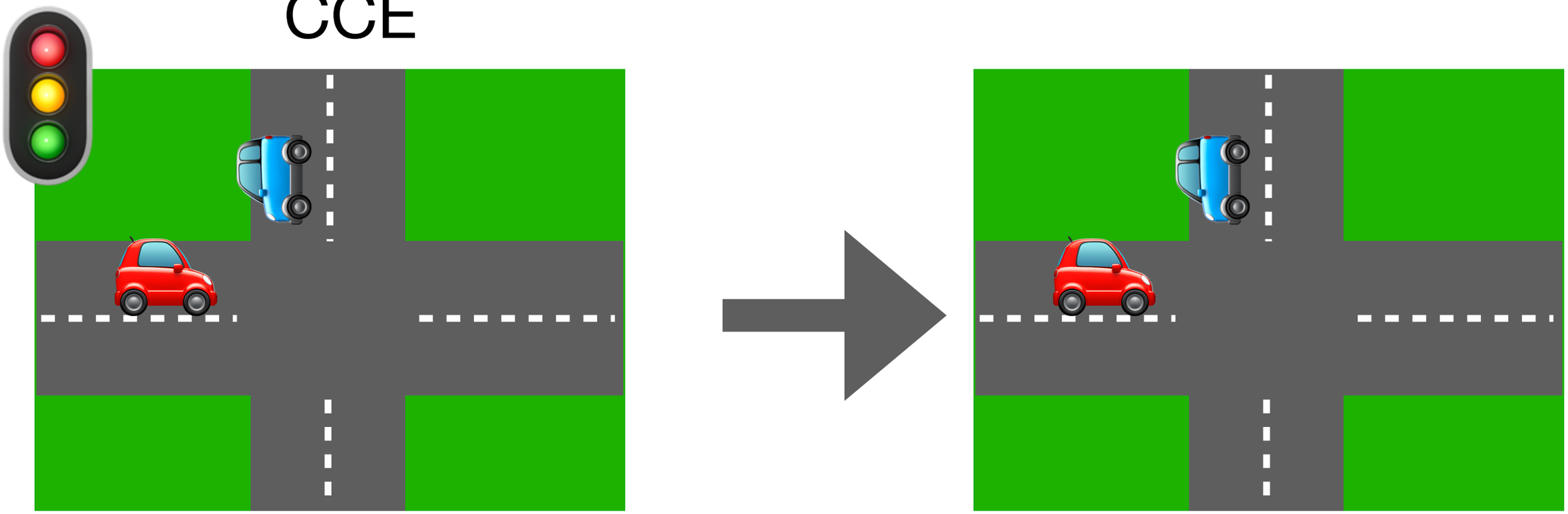


	Go	Stop
Go	0	0.5
Stop	0.5	0

	0.5 Go	0.5 Stop
0.5 Go	0.25	0.25
0.5 Stop	0.25	0.25

What If Players Lose the Ability to Correlate?

CCE

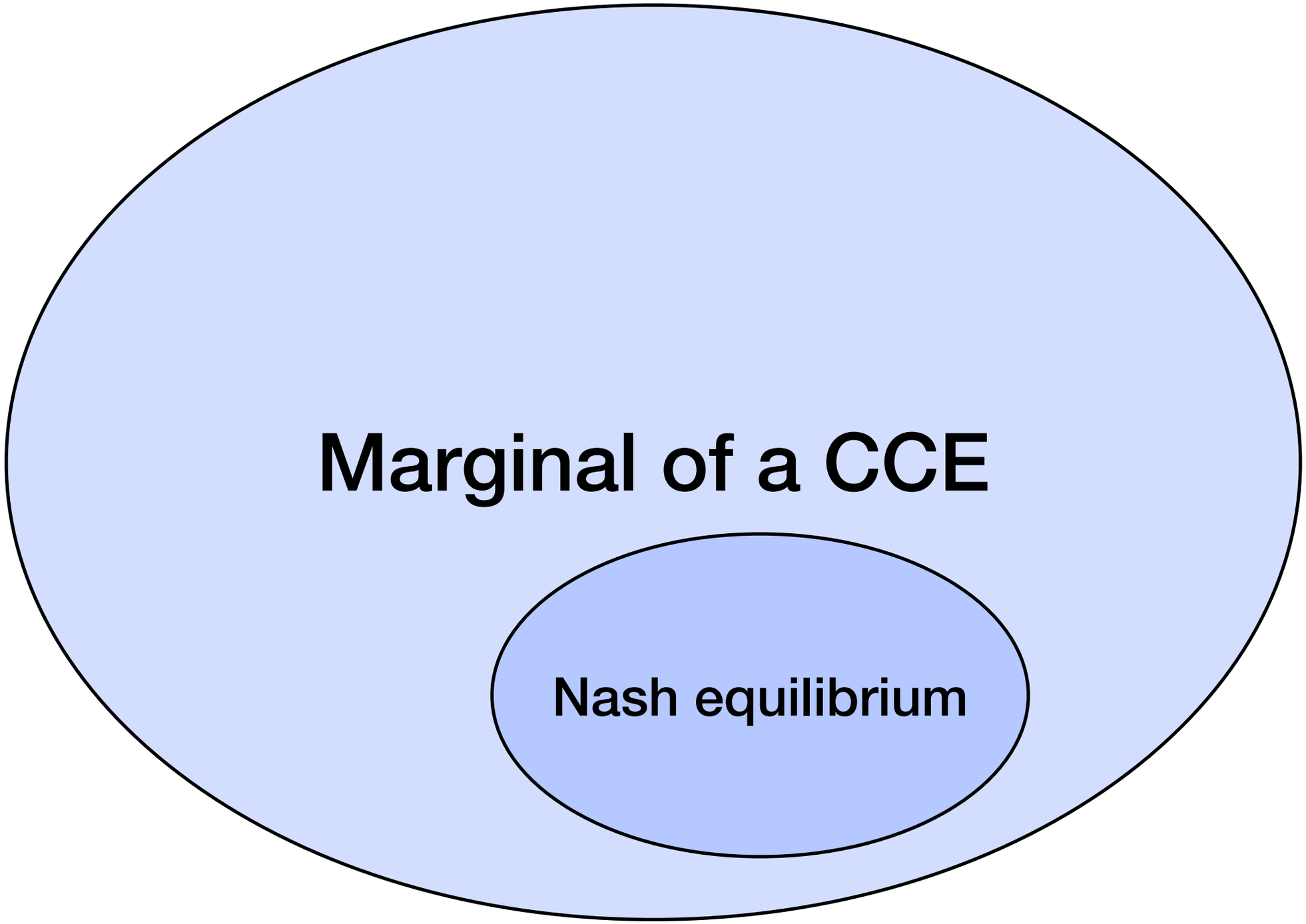


$$\mu \in \Delta(P)$$

CCE

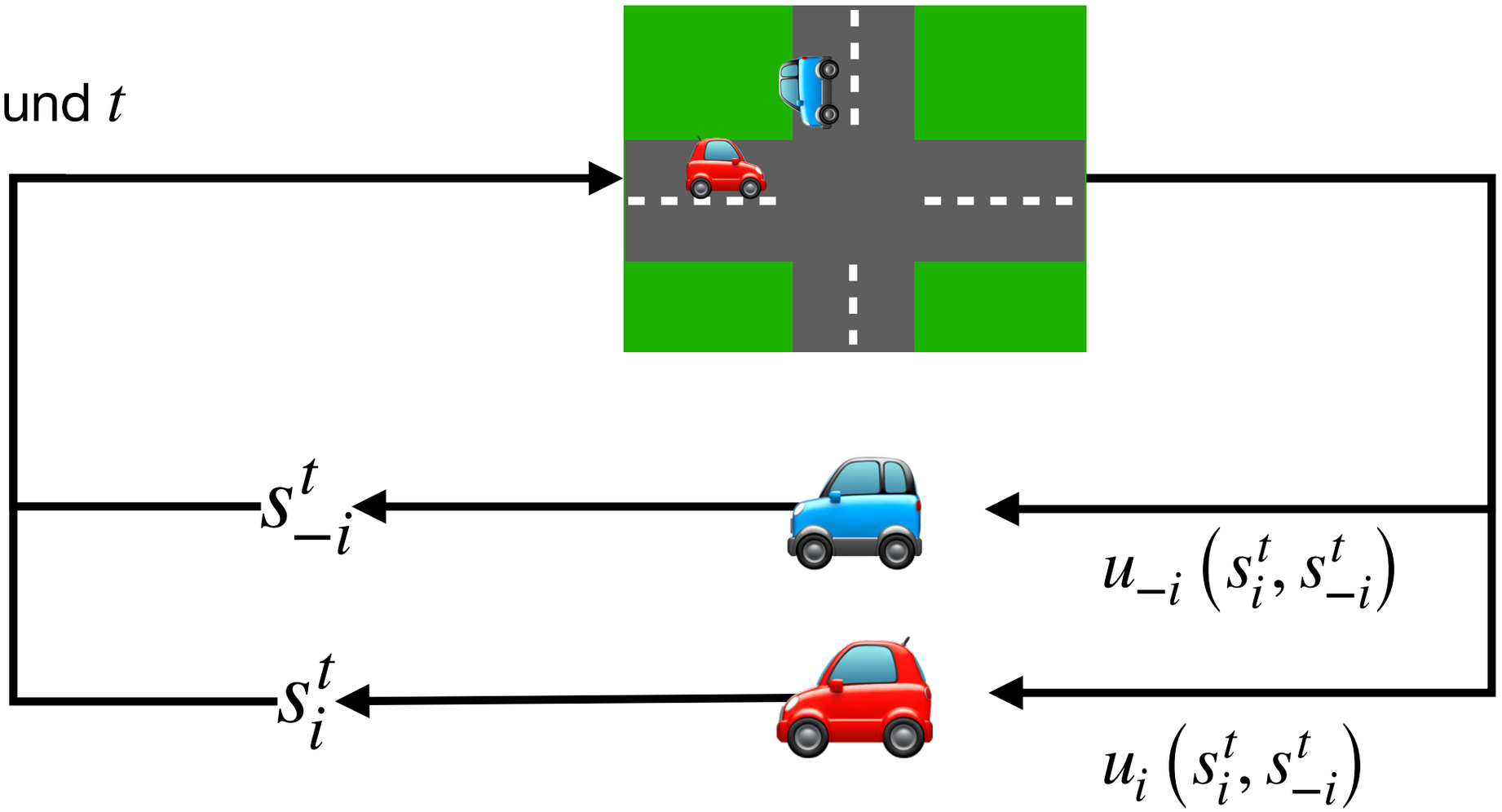
$$S^{\mu}$$

Marginal
strategy profile

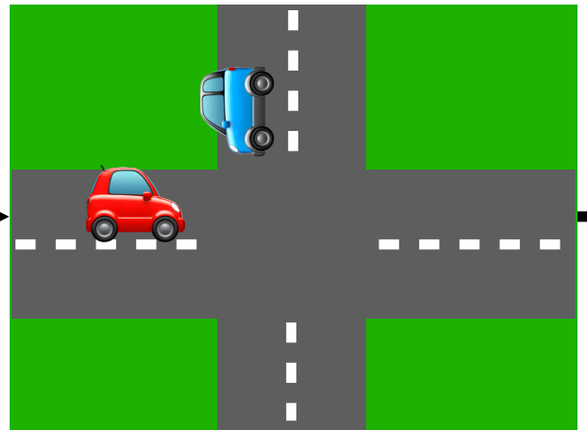

Marginal of a CCE
Nash equilibrium

Self-Play & No-Regret Learning

round t



round t

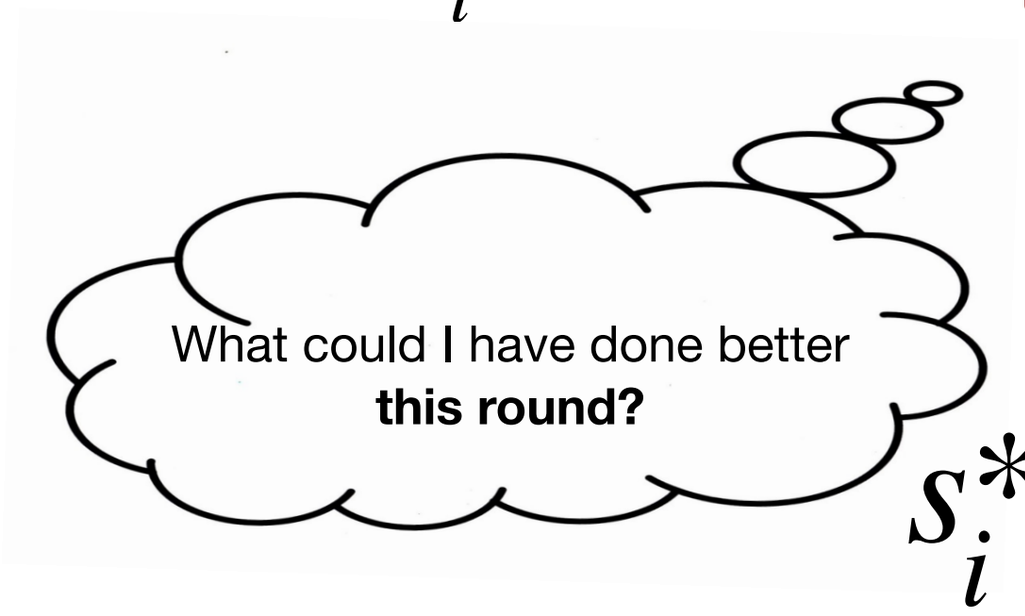


s_{-i}^t

$u_{-i}\left(s_i^t, s_{-i}^t\right)$

s_i^t

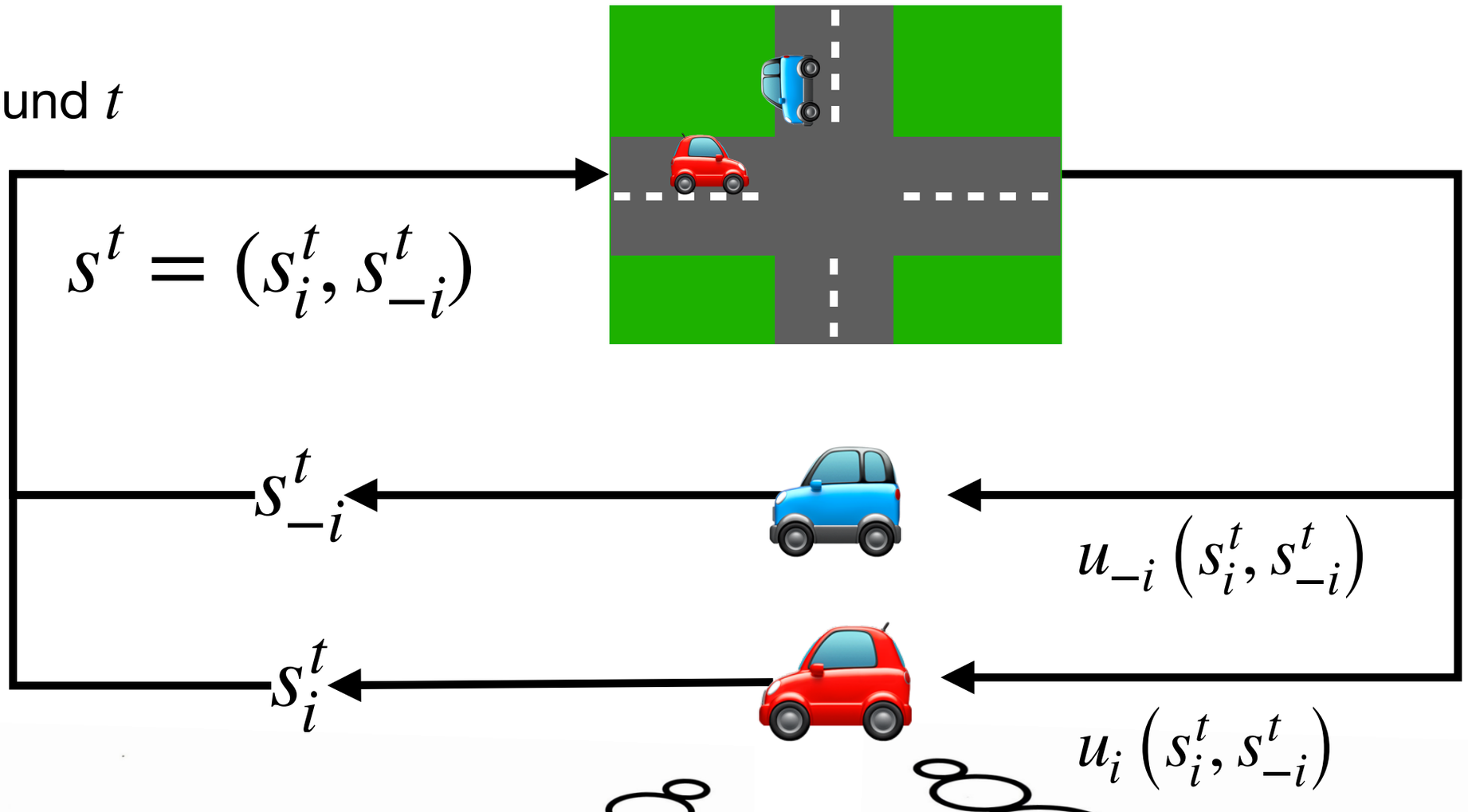
$u_i\left(s_i^t, s_{-i}^t\right)$



s_i^*

$u_i\left(s_i^t, s_{-i}^t\right)$

round t



What could I have done better **this round?**

What could I have done better in all **past rounds**?

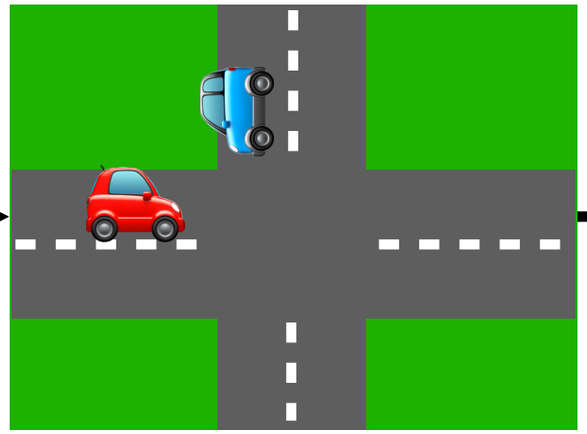
Regret

$$R_i(s_i^*, s^t) \doteq u_i\left(s_i^*, s_{-i}^t\right) - u_i\left(s_i^t, s_{-i}^t\right)$$

Average Regret

$$\frac{1}{T}\sum_{t=1}^{T} R_i(s_i^*, s^t)$$

round t



$s^t = (s_i^t, s_{-i}^t)$

s_{-i}^t

$u_{-i}\left(s_i^t, s_{-i}^t\right)$

s_i^t

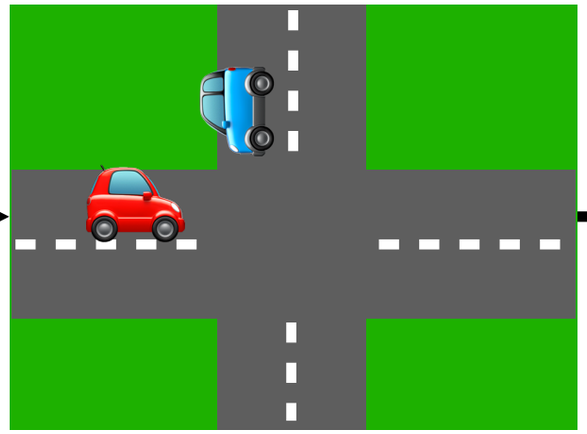
$u_i\left(s_i^t, s_{-i}^t\right)$

$\mathscr{A}$

no-regret algorithm

Goal: Average regret $\to 0$

round t



$$s^t = (s_i^t, s_{-i}^t)$$

s_{-i}^t

$\mathscr{A}$

$u_{-i}\left(s_i^t, s_{-i}^t\right)$

s_i^t

$\mathscr{A}$

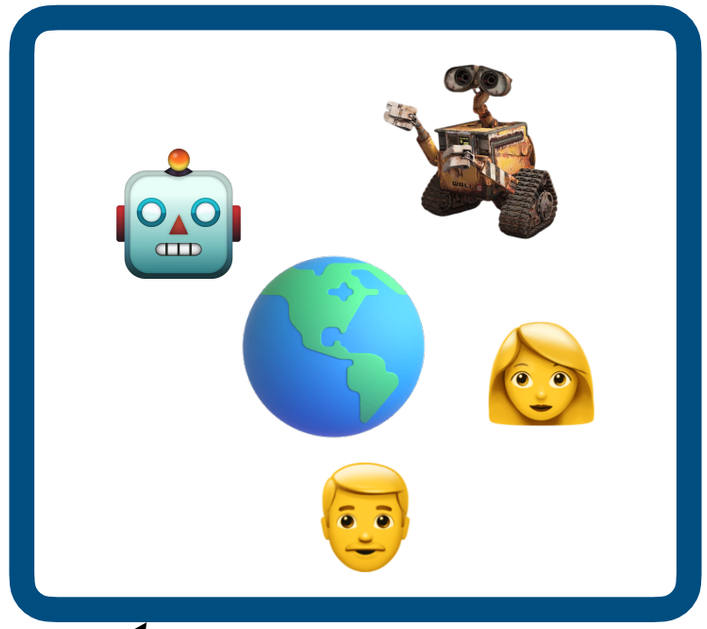
$u_i\left(s_i^t, s_{-i}^t\right)$

no-regret self-play

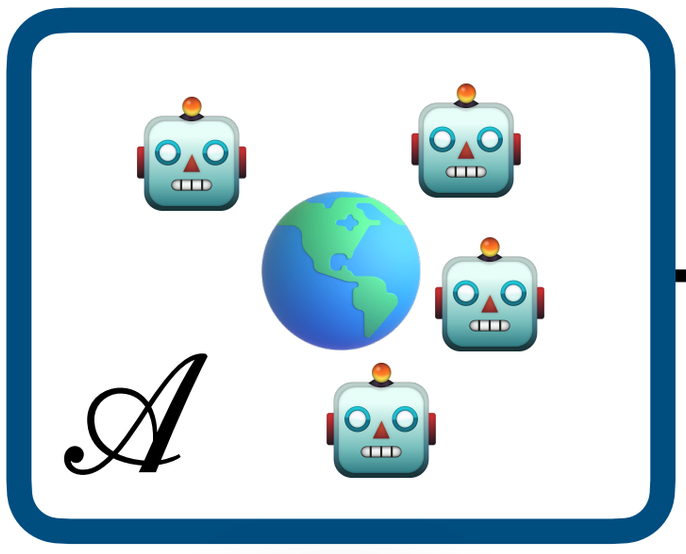
Marginal strategy of a CCE

$$\frac{1}{T}\sum_{t=1}^{T} s_i^t \to s_i^\mu$$

Online Interaction



Offline Self-Play



s^{μ}

s_i^{μ}

We characterize the strategy produced by self-play as the marginal strategy of a CCE

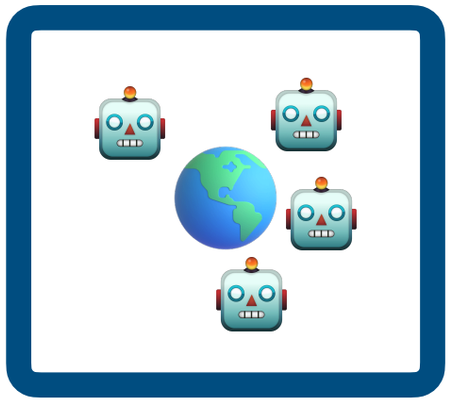
What Do We Mean By “Good” Strategies?

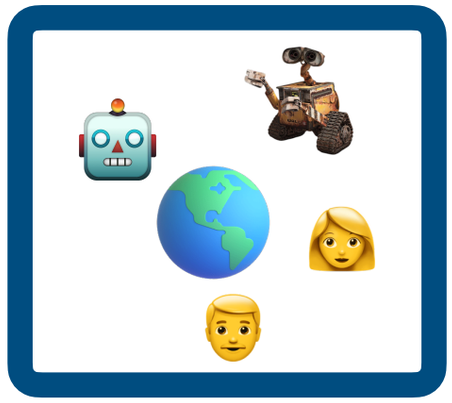
What Do We Mean By “Good” Strategies?

What Do We Mean By “Good” Strategies?

Self-Play Wishlist

- ☐ Low **vulnerability**





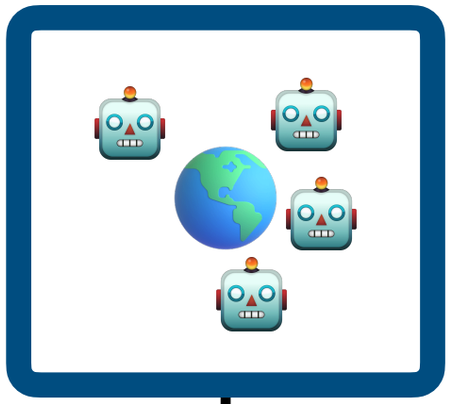
$$u_i(s_i^{\mu}, s_{-i}^{\mu}) - \min_{s_{-i}} u_i(s_i^{\mu}, s_{-i}) \leq \epsilon$$

What Do We Mean By “Good” Strategies?

Self-Play Wishlist

- ☐ Low **vulnerability**
- ☐ Similar **values**

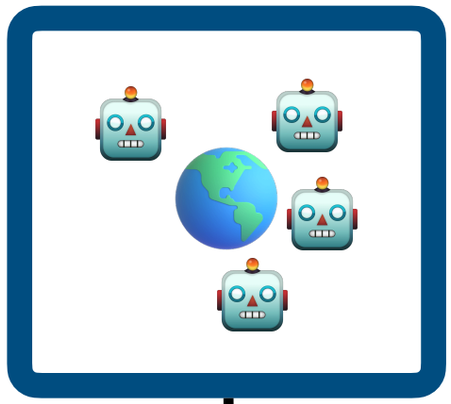
Random seed = 0



s^1

$u_i(s^1)$

Random seed = 32



s^2

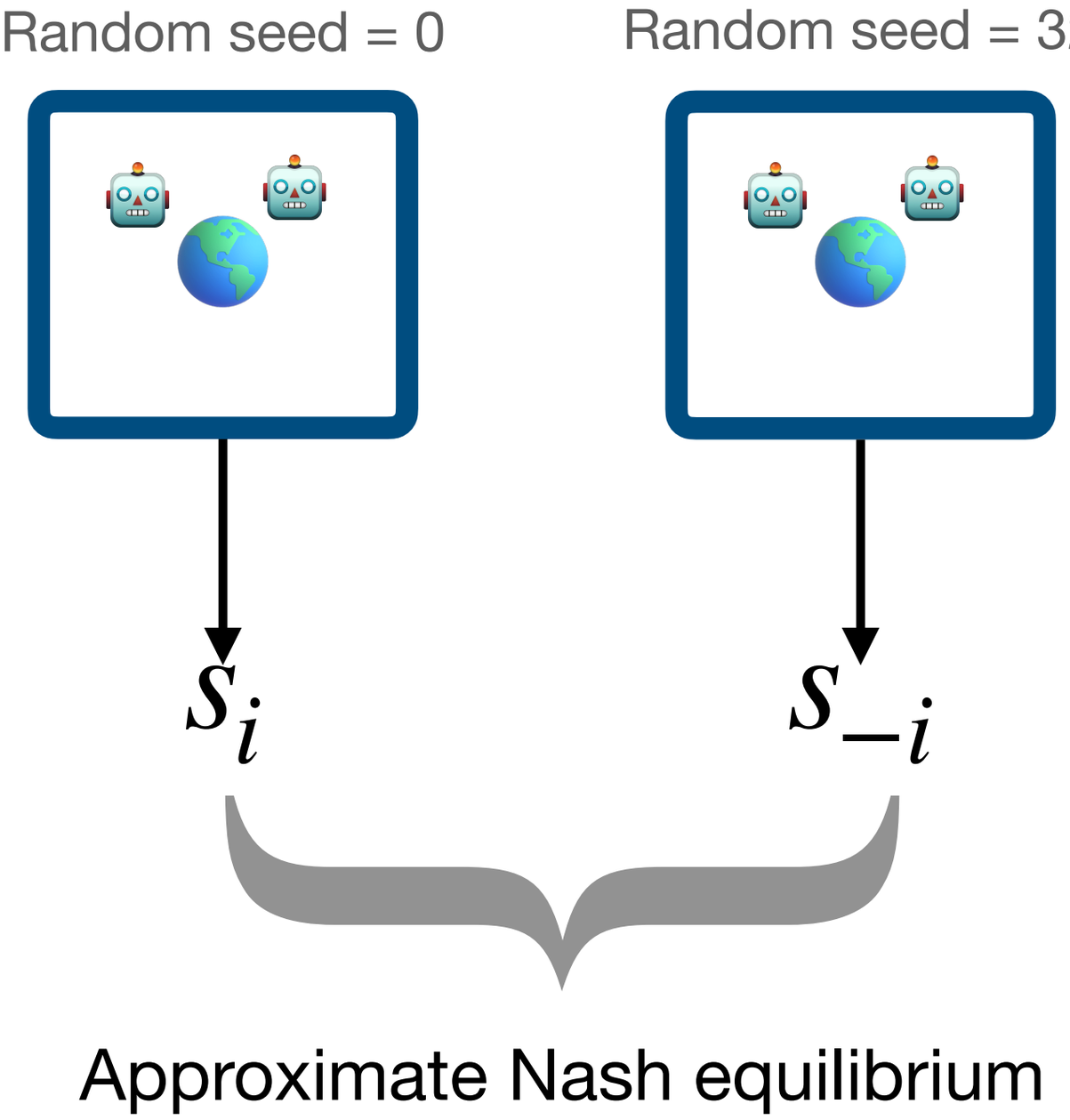
$u_i(s^2)$

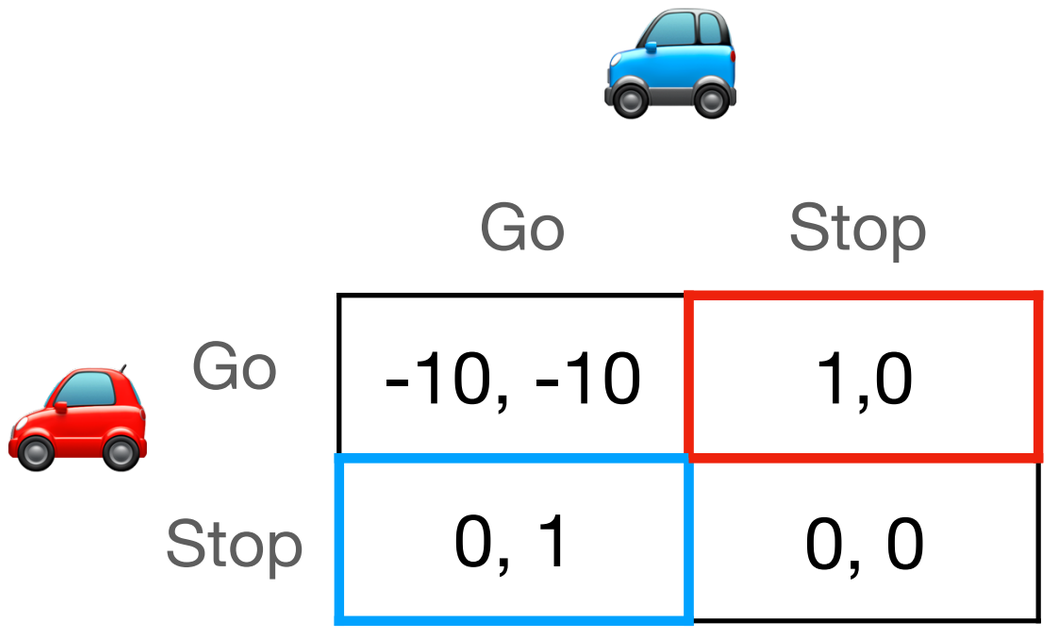
$\approx$

What Do We Mean By “Good” Strategies?

Self-Play Wishlist

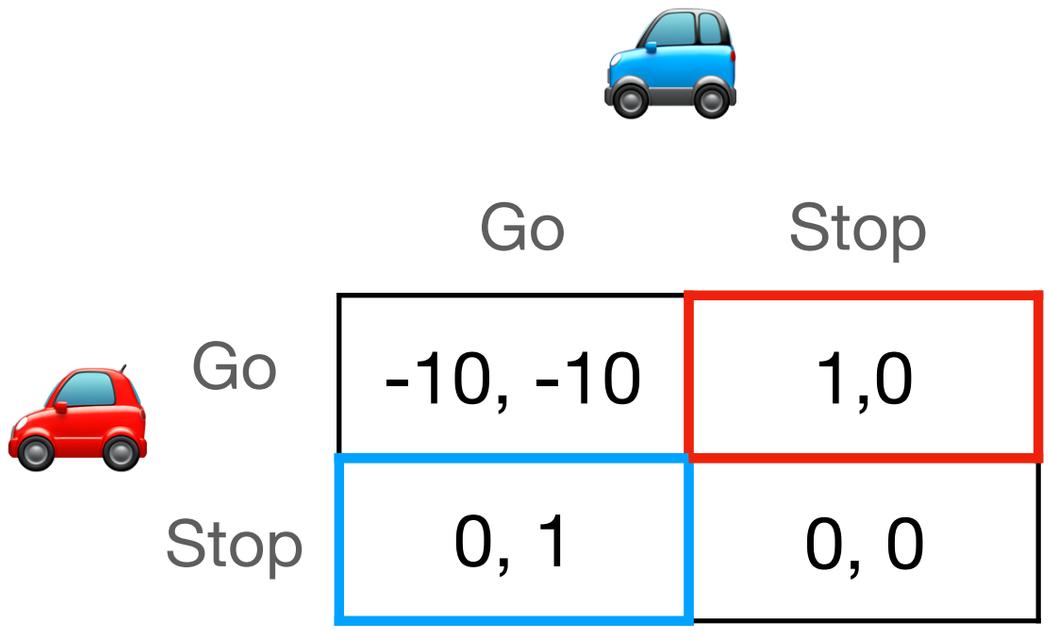
- ☐ Low **vulnerability**
- ☐ Similar **values**
- ☐ Nearly **Exchangeable**





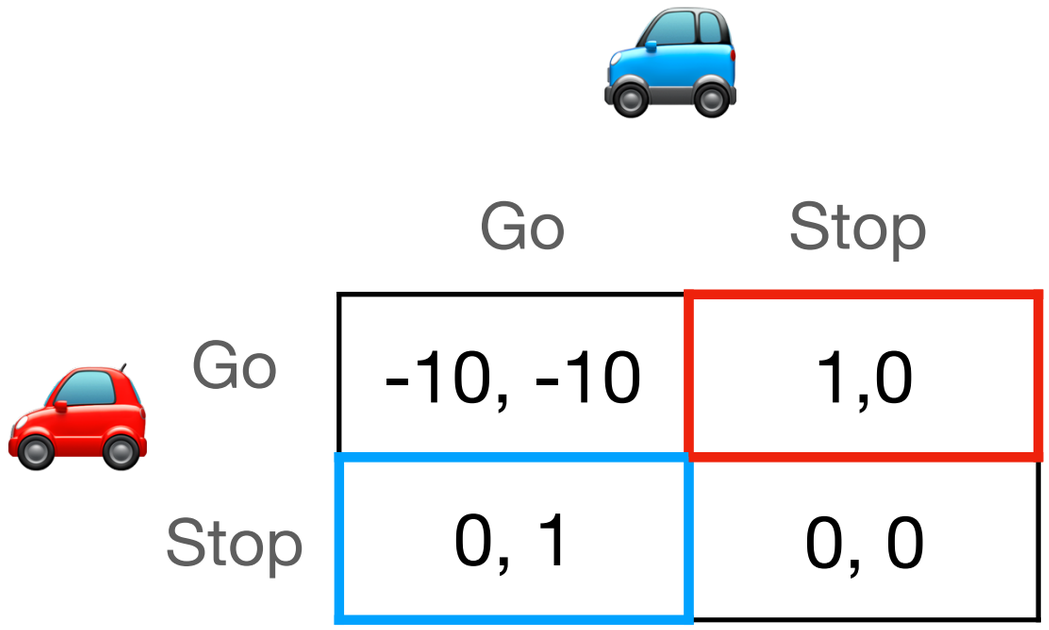
Self-Play Wishlist

- ☐ Low **vulnerability**
- ☐ Similar **values**
- ☐ Nearly **Exchangeable**



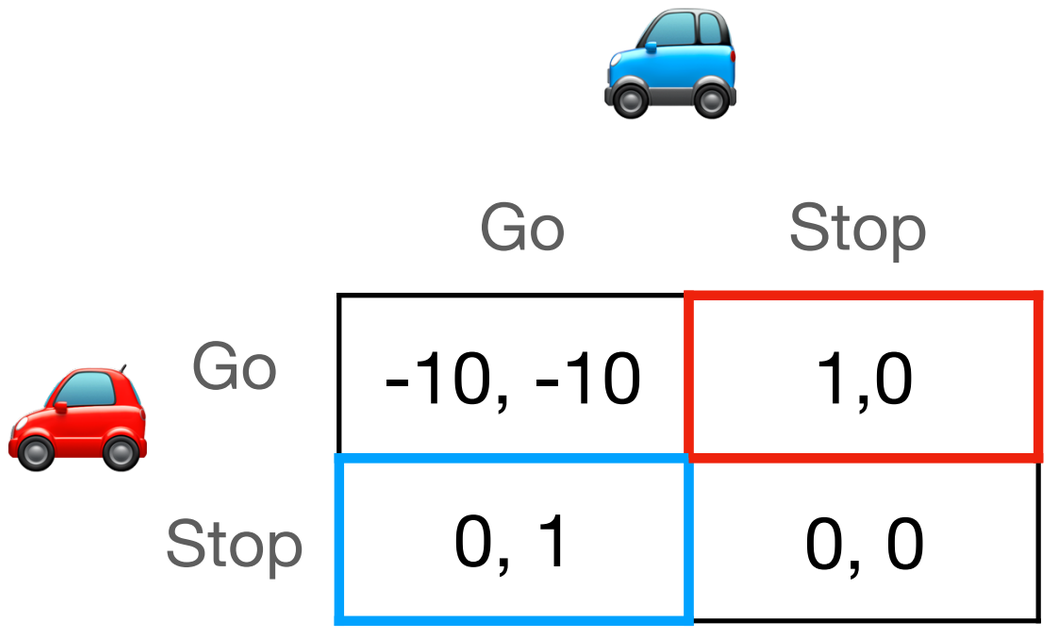
Self-Play Wishlist

- ☐ ~~Low **vulnerability**~~
- ☐ Similar **values**
- ☐ Nearly **Exchangeable**



Self-Play Wishlist

- ☐ ~~Low **vulnerability**~~
- ☐ ~~Similar **values**~~
- ☐ Nearly **Exchangeable**



Self-Play Wishlist

- ☐ ~~Low **vulnerability**~~
- ☐ ~~Similar **values**~~
- ☐ ~~Nearly **Exchangeable**~~

When Does Self-Play Have Guarantees?

Two Player Constant-Sum Games

2 player

- Self-play will produce an approximate Nash equilibrium

Self-Play Wishlist

☑ Low **vulnerability**

☑ Similar **values**

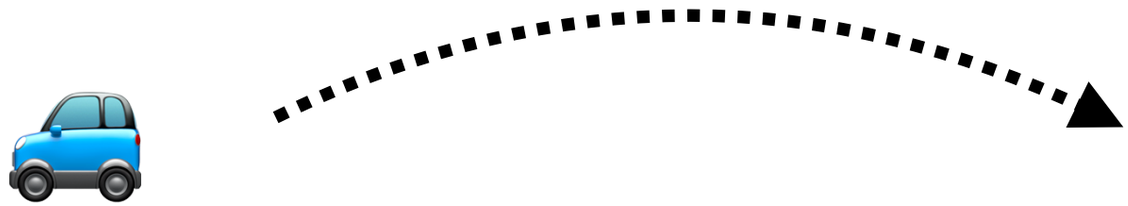
☑ Nearly **Exchangeable**

What About Constant-Sum Multi-player Games?

	Go	Stop
Go	-10, -10	1,0
Stop	0, 1	0, 0

	Go	Stop
Go	-10, -10, **20**	1,0, **-1**
Stop	0, 1, **-1**	0, 0, **0**

What About Constant-Sum Multi-player Games?



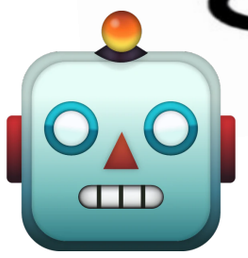
	Go	Stop
Go	-10, -10	1,0
Stop	0, 1	0, 0

	Go	Stop
Go	-10, -10, **20**	1,0, **-1**
Stop	0, 1, **-1**	0, 0, **0**

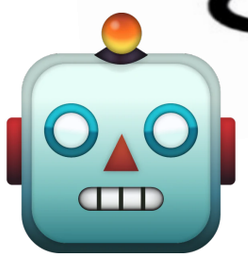
Self-Play Wishlist

- ☐ ~~Low **vulnerability**~~
- ☐ ~~Similar **values**~~
- ☐ ~~Nearly **Exchangeable**~~

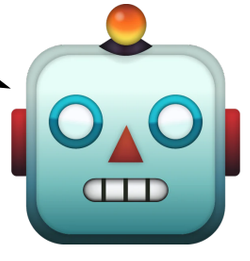
What are properties of games where self-play produce **good** strategies?



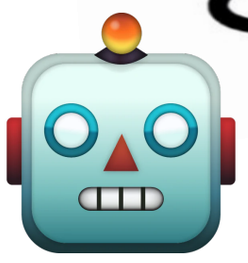
What are properties of games where self-play produce **good** strategies?



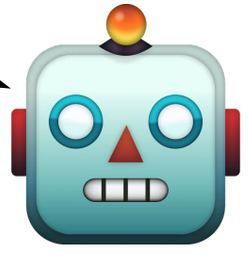
Two-player constant-sum games…



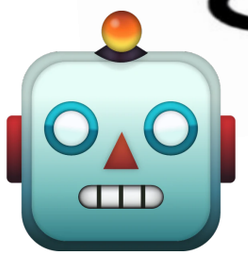
What are properties of games where self-play produce **good** strategies?



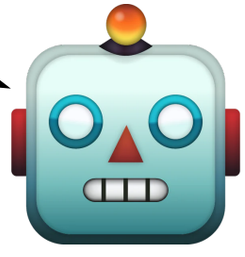
But not necessarily multi-player constant-sum games



What are properties of games where self-play produce **good** strategies?



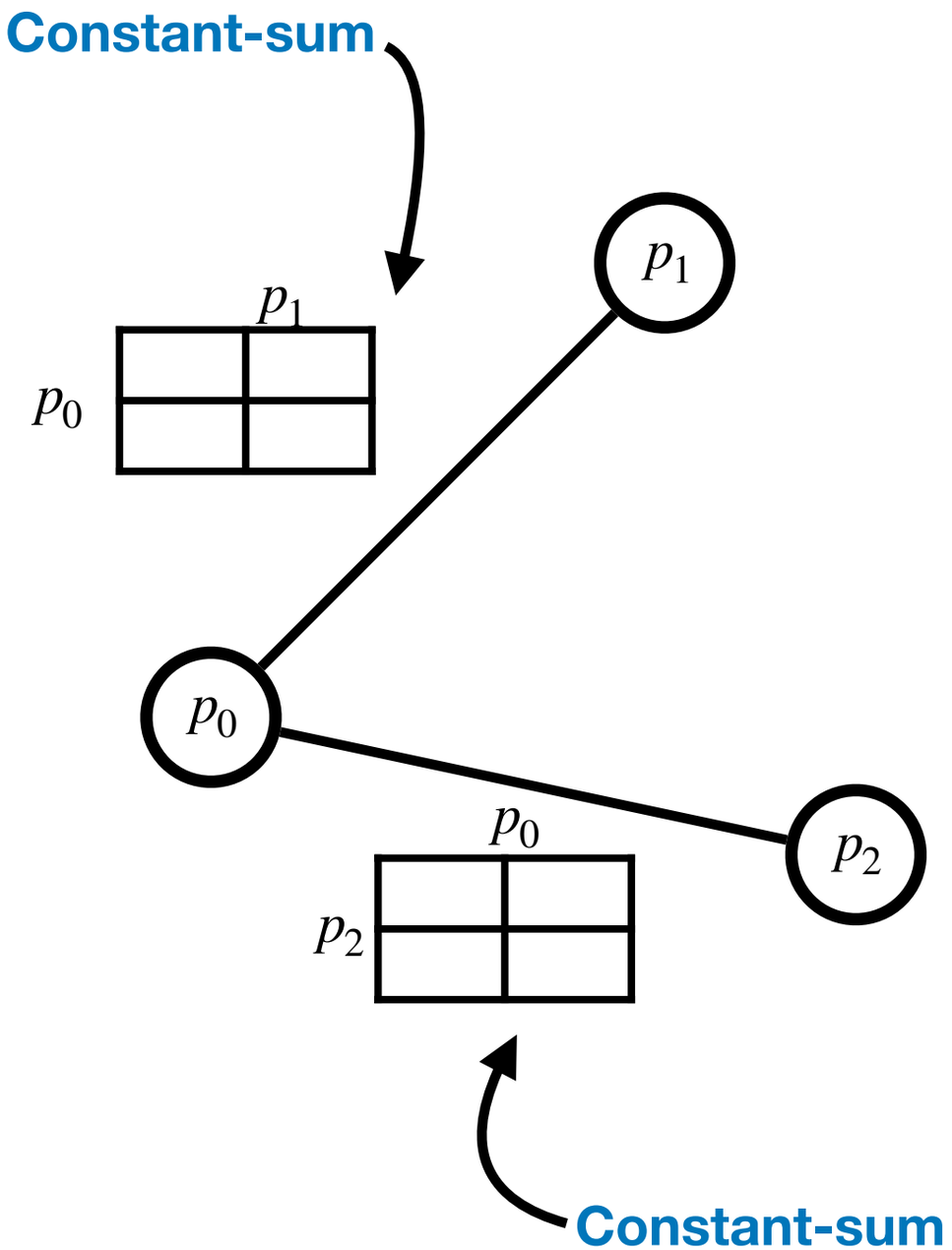
But what else?



There are multi-player games that are **structurally similar** to two-player constant-sum games

Constant-Sum Polymatrix (CSP) Games

Bregman & Fokin, 1987; Cai et al., 2011



Overall utility = **sum** of subgame utilities

Approximate CSP Games

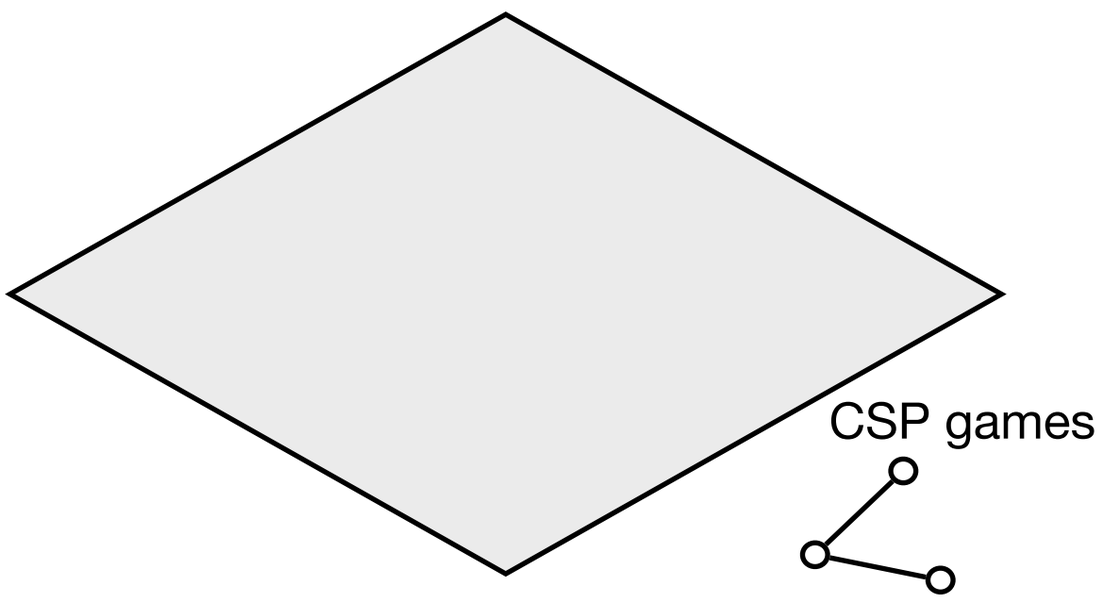
- Not every game has an **exact** CSP decomposition
- But every game can be **approximated** by a CSP game!

$$G$$

Approximate CSP Games

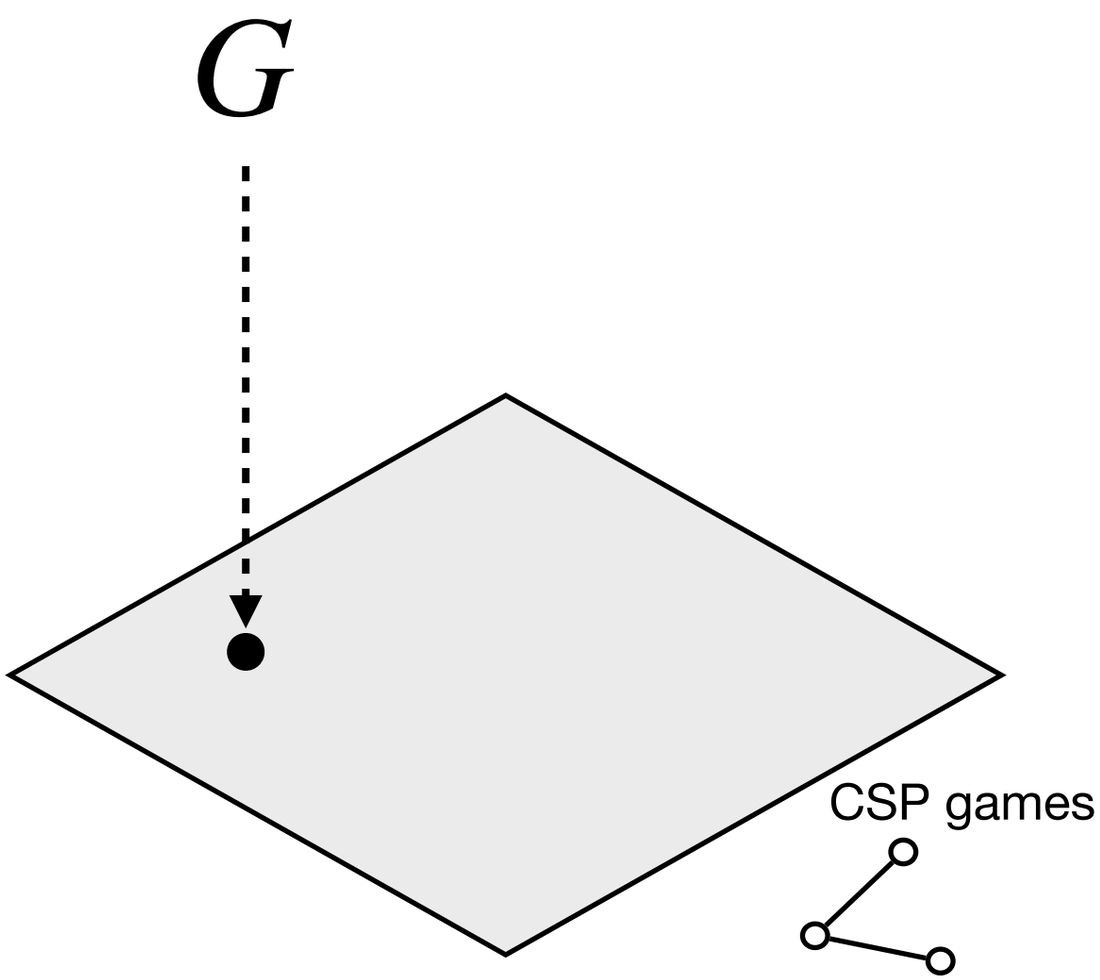
- Not every game has an **exact** CSP decomposition
- But every game can be **approximated** by a CSP game!

G



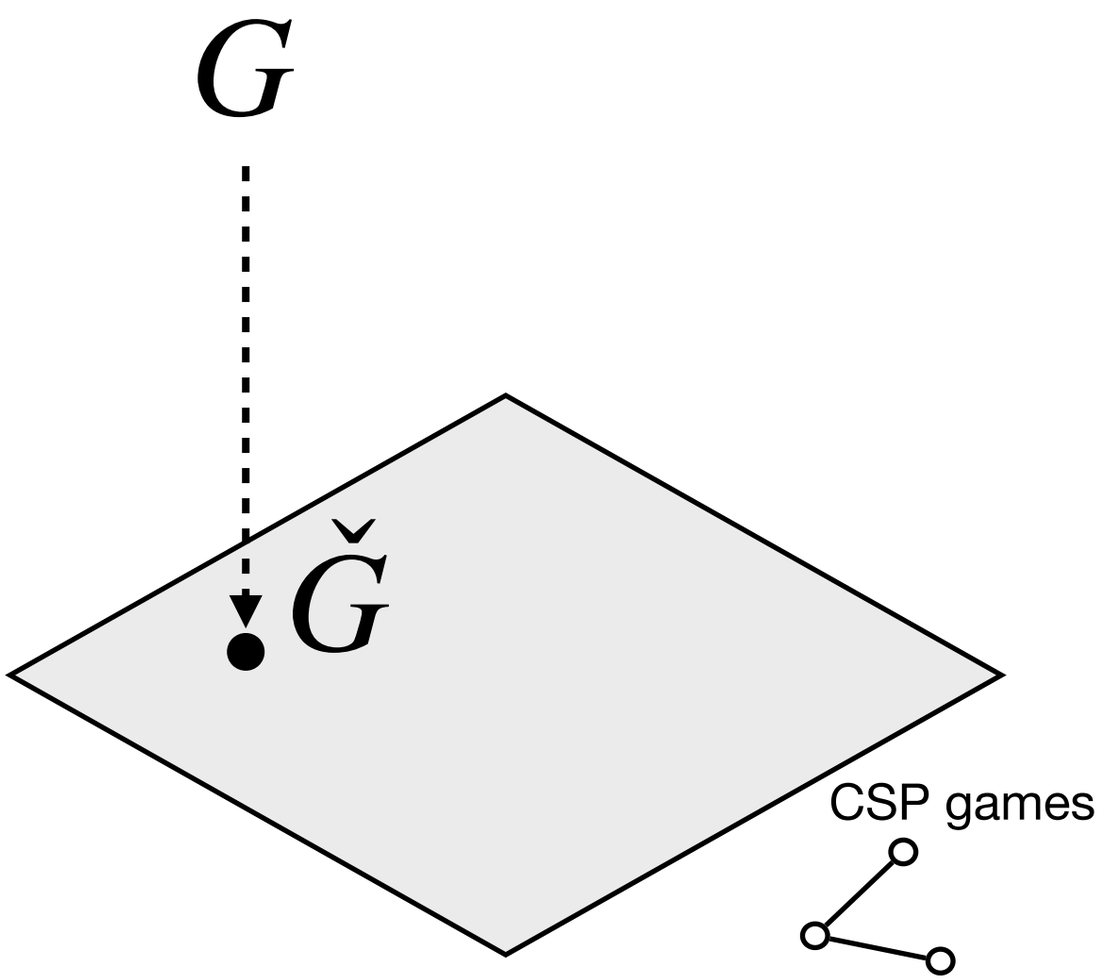
Approximate CSP Games

- Not every game has an **exact** CSP decomposition
- But every game can be **approximated** by a CSP game!



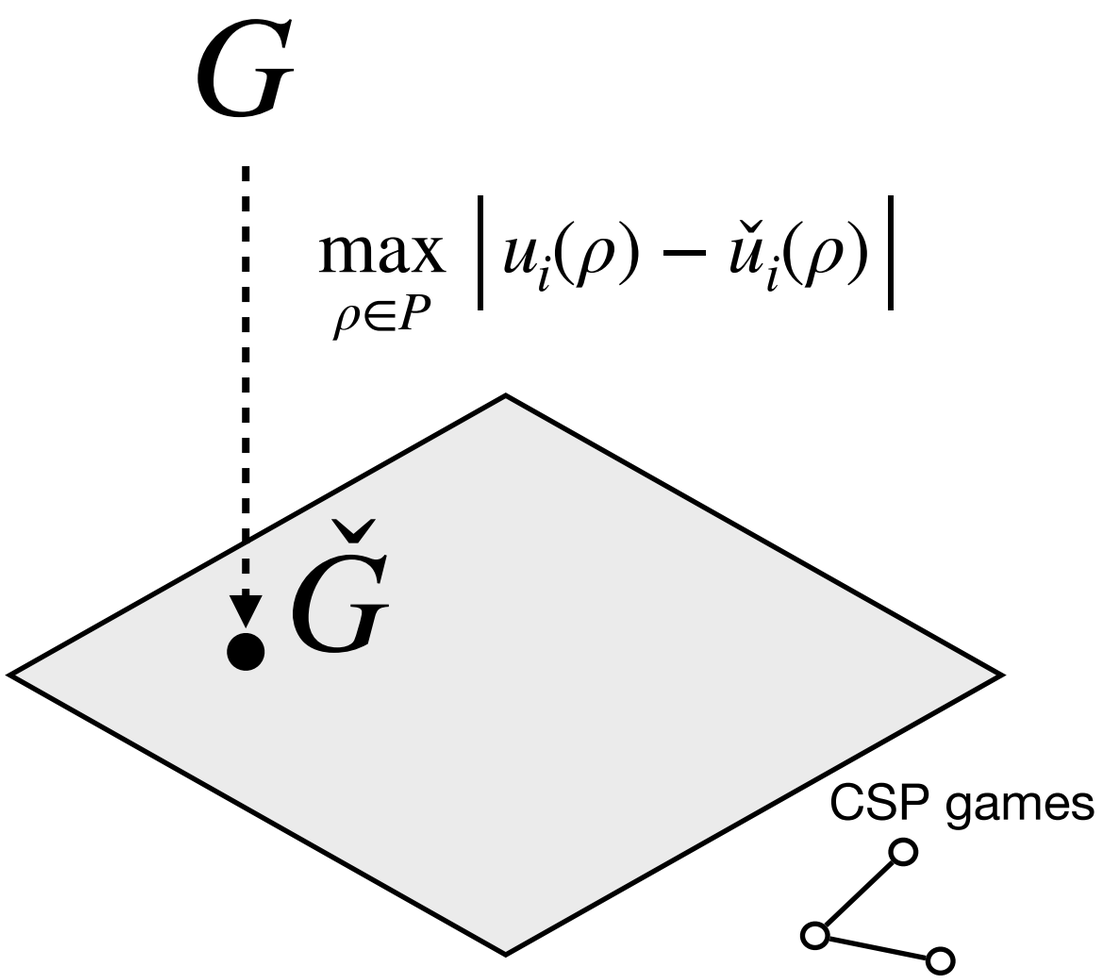
Approximate CSP Games

- Not every game has an **exact** CSP decomposition
- But every game can be **approximated** by a CSP game!



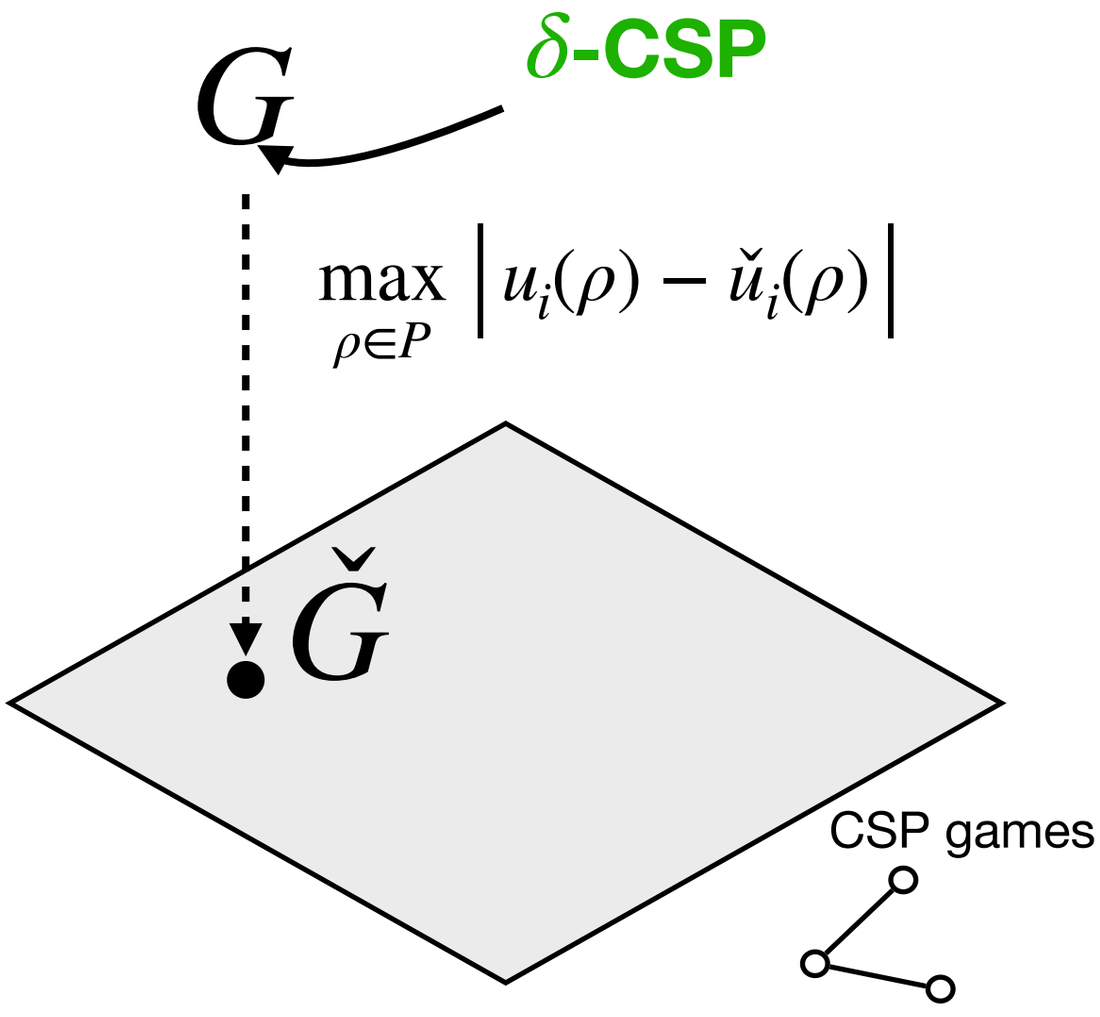
Approximate CSP Games

- Not every game has an **exact** CSP decomposition
- But every game can be **approximated** by a CSP game!



Approximate CSP Games

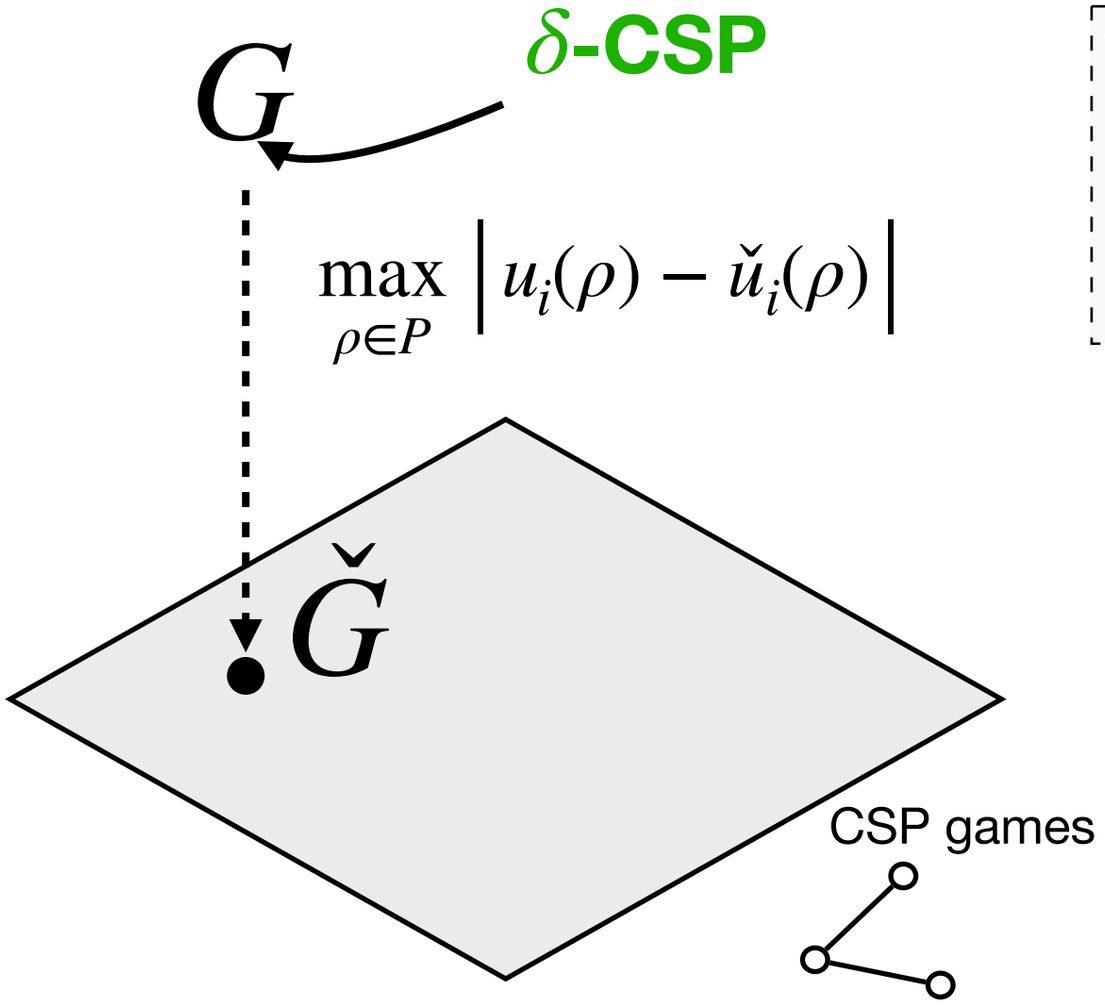
- Not every game has an **exact** CSP decomposition
- But every game can be **approximated** by a CSP game!



Polynomial time!

Approximate CSP Games

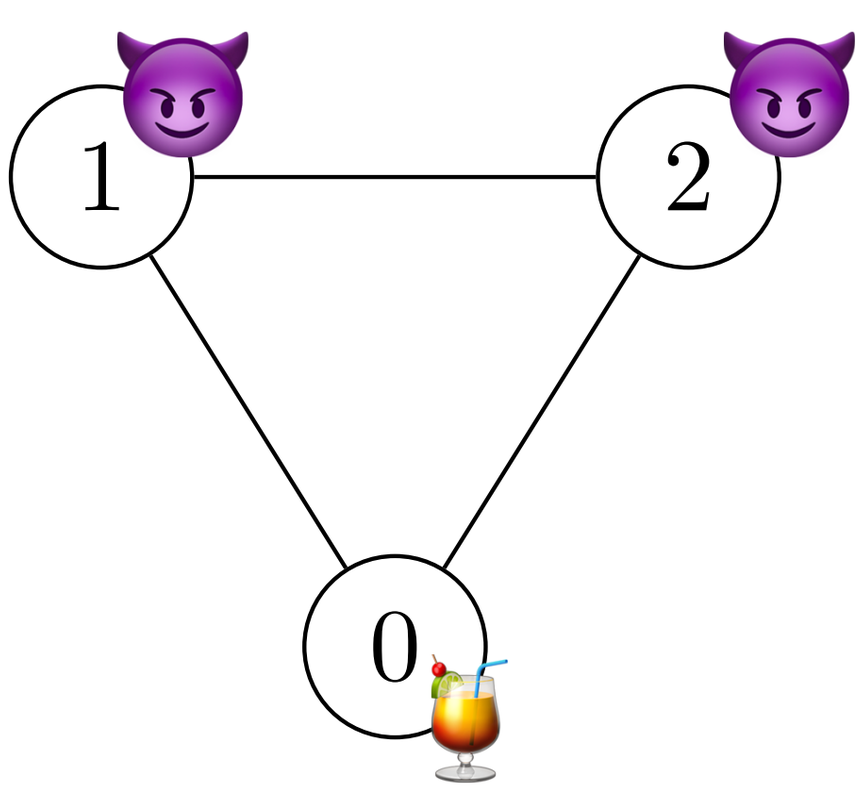
- Not every game has an **exact** CSP decomposition
- But every game can be **approximated** by a CSP game!



Proposition. Self-play will produce a $2(n+1)\delta$-Nash equilibrium in δ-CSP games.

Counterexample

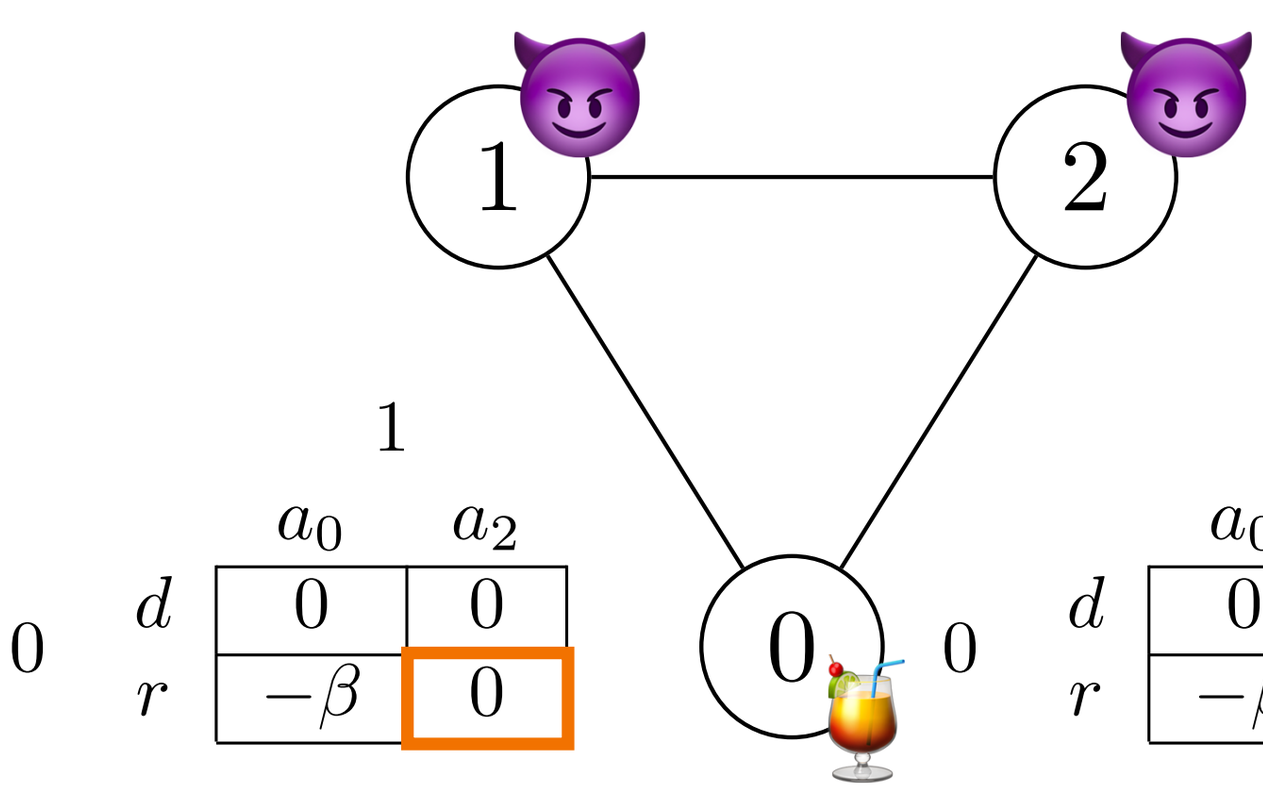
Offense-Defense



Counterexample

Offense-Defense

$$
\begin{array}{cc|cc}
 & 1 & a_0 & a_2 \\
2 & a_0 & 0 & -\beta \\
 & a_1 & \beta & 0
\end{array}
$$



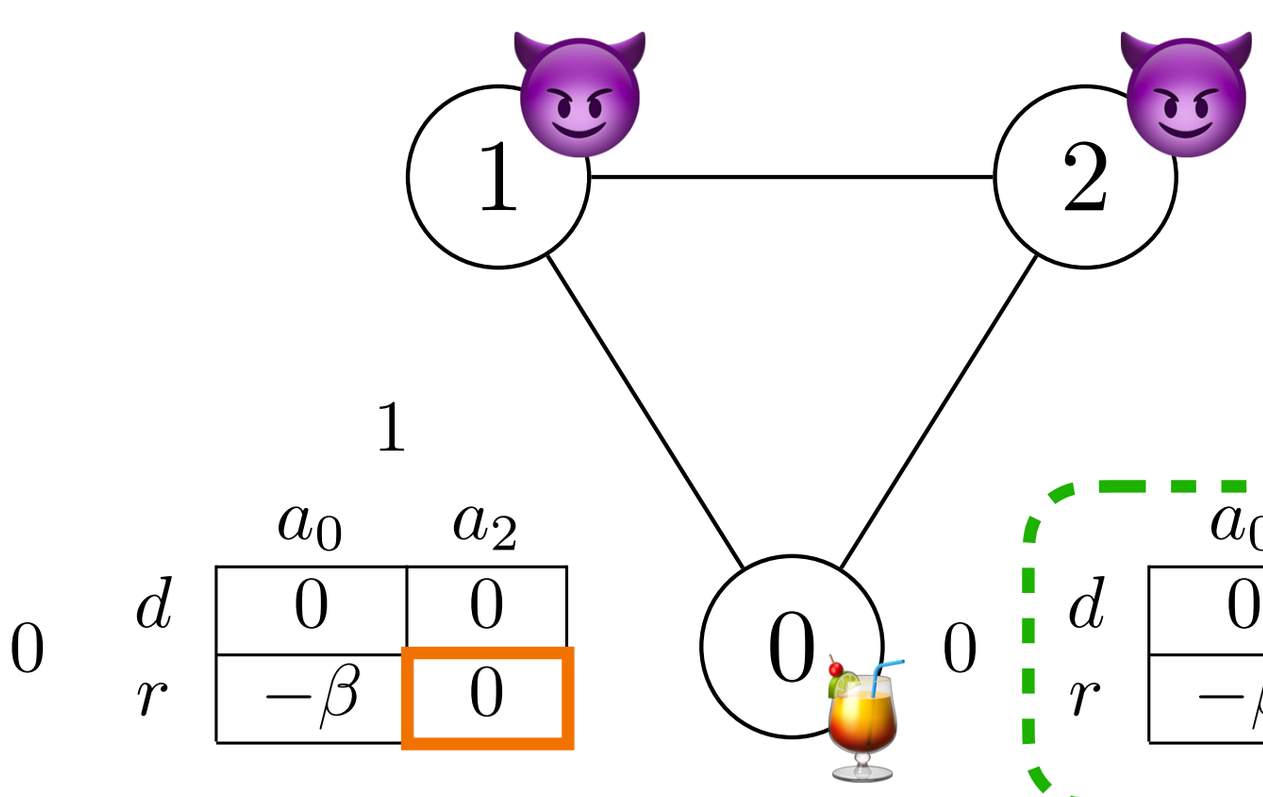
$$
\begin{array}{cc|cc}
 & 1 & a_0 & a_2 \\
0 & d & 0 & 0 \\
 & r & -\beta & 0
\end{array}
$$

$$
\begin{array}{cc|cc}
 & 2 & a_0 & a_1 \\
0 & d & 0 & 0 \\
 & r & -\beta & 0
\end{array}
$$

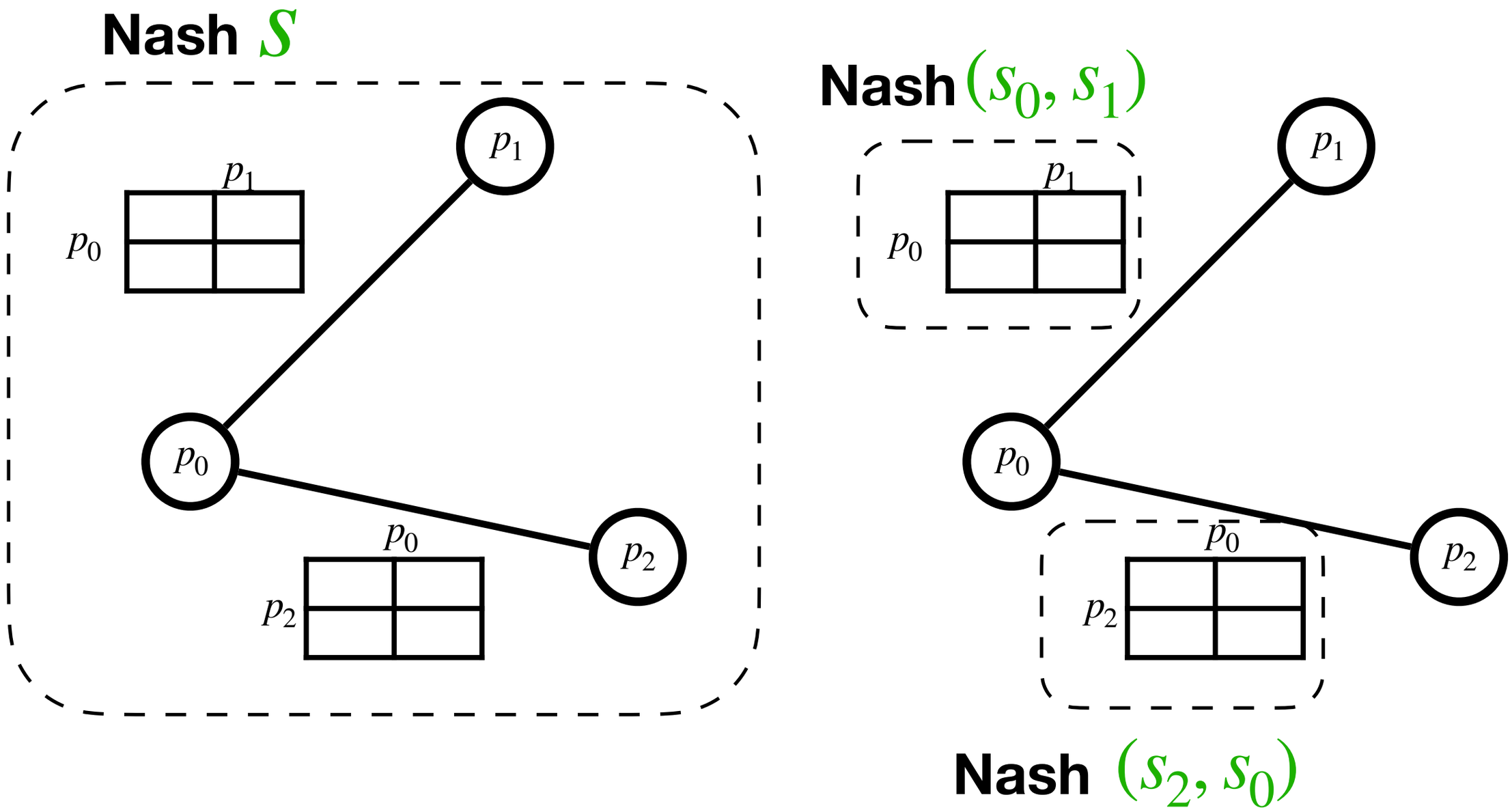
Counterexample

Offense-Defense

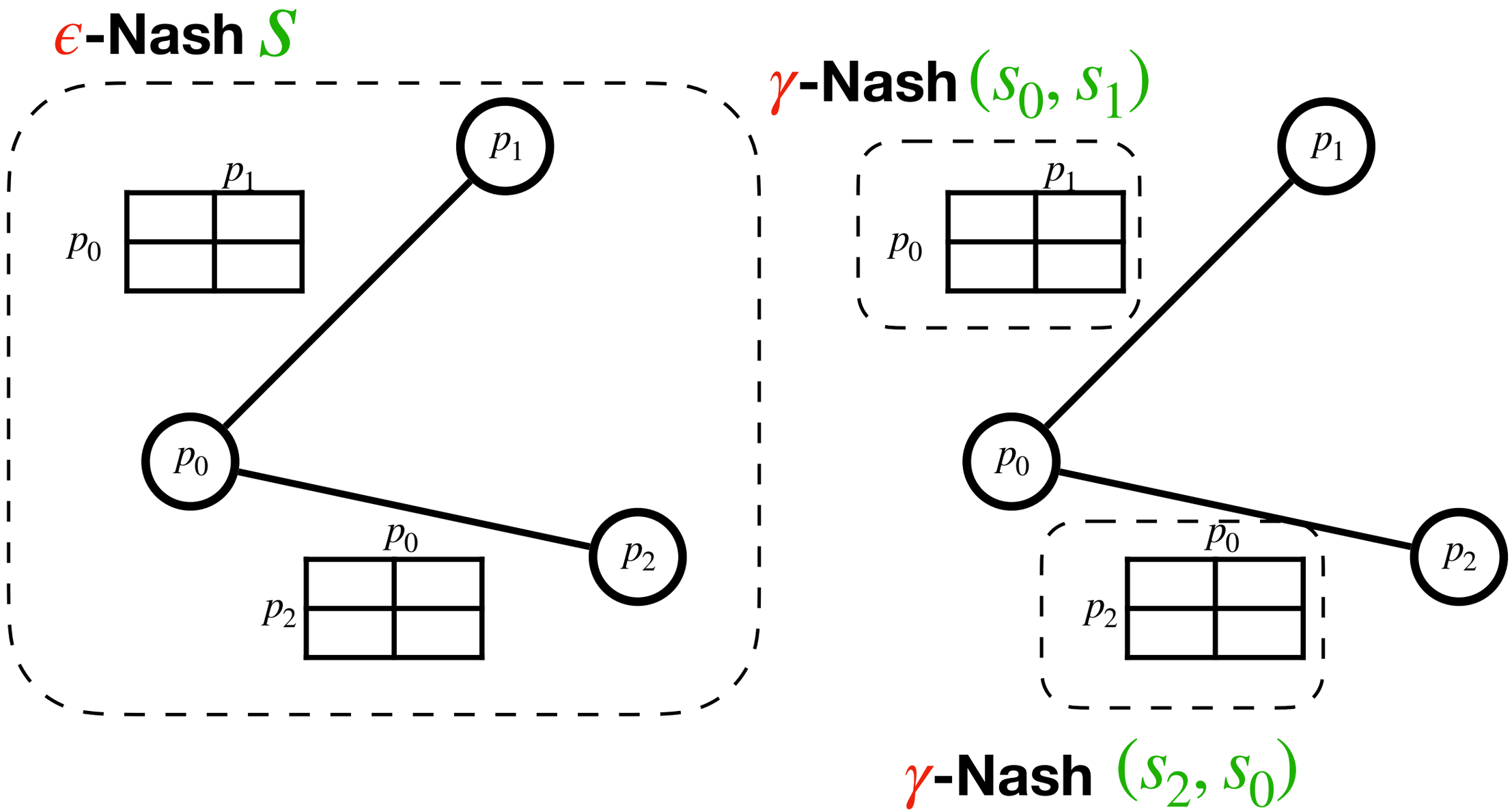
$$
\begin{array}{c|c|cc}
 & & \multicolumn{2}{c}{1} \\
 & & a_0 & a_2 \\
\hline
2 & a_0 & 0 & -\beta \\
\hline
 & a_1 & \beta & 0 \\
\hline
\end{array}
$$



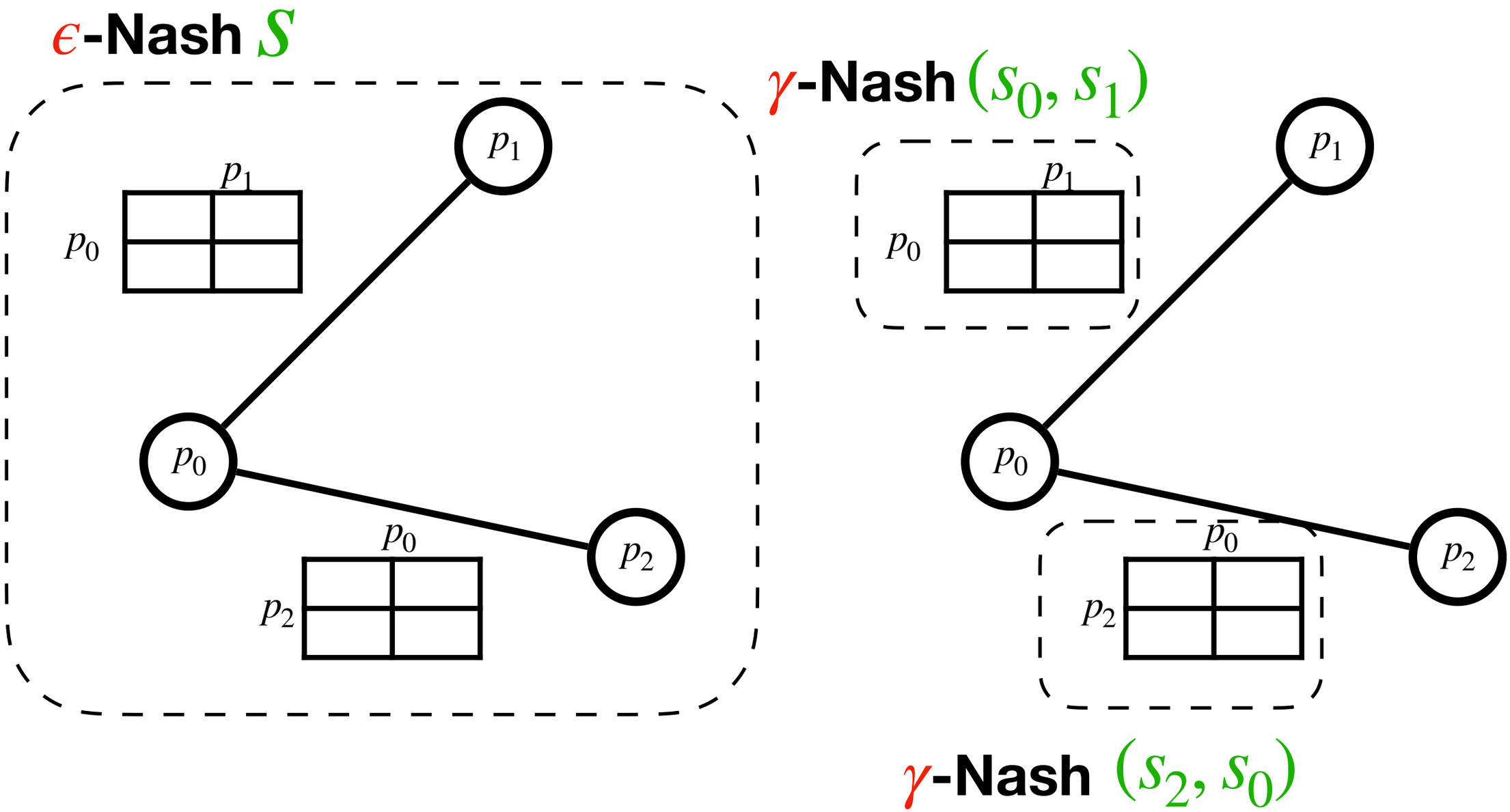
Subgame Stability



Approximate Subgame Stability



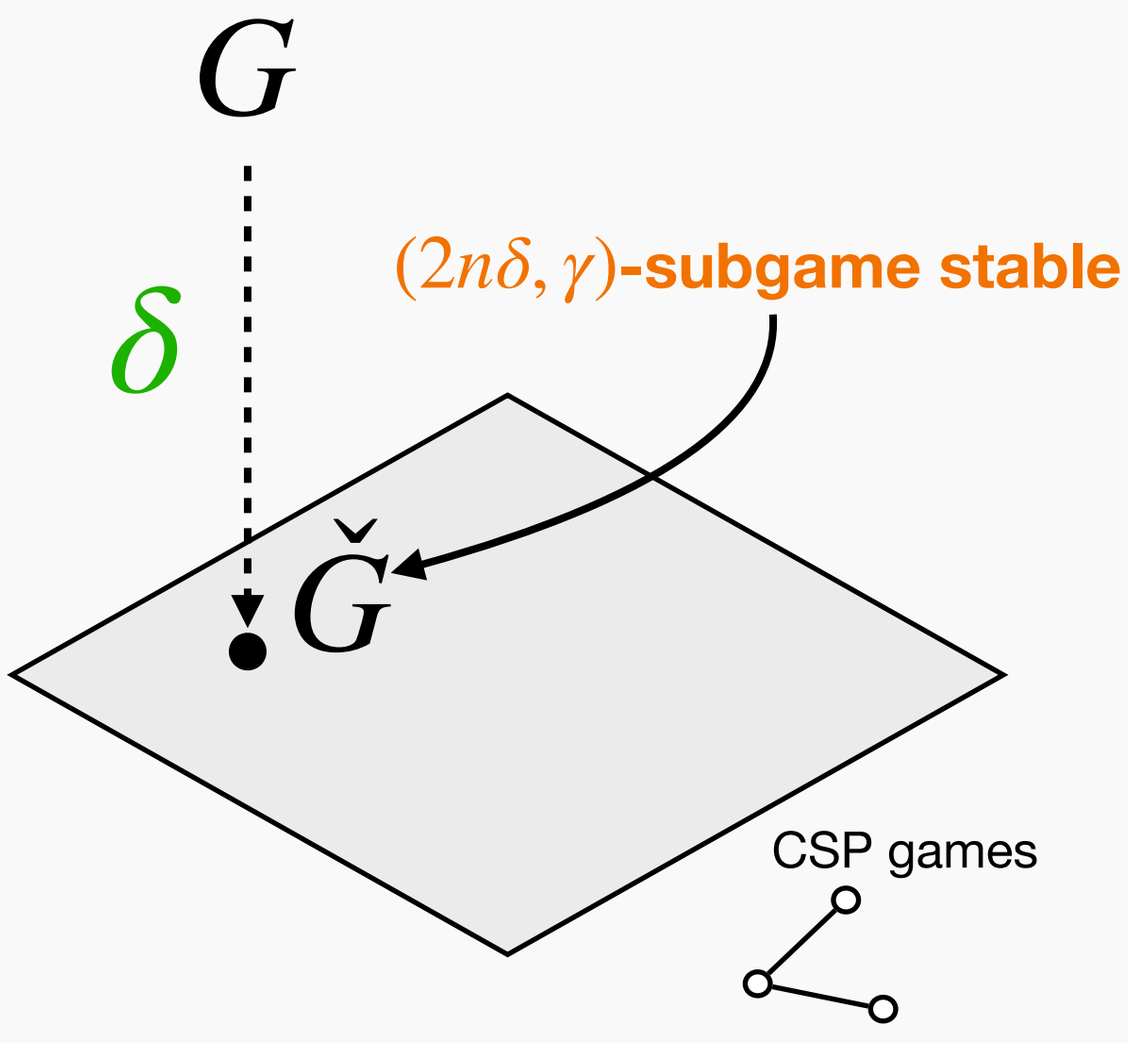
(ϵ, γ)-Subgame Stability



Main Result

Guarantees For Self-Play in Multi-Player Games

Theorem. If

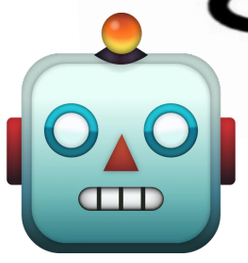


Then,

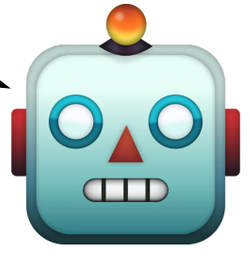
Self-Play Wishlist

- ☑ Low **vulnerability**
- ☑ Similar **values**
- ☑ Nearly **Exchangeable**

What are properties of games where self-play produces **good** strategies?

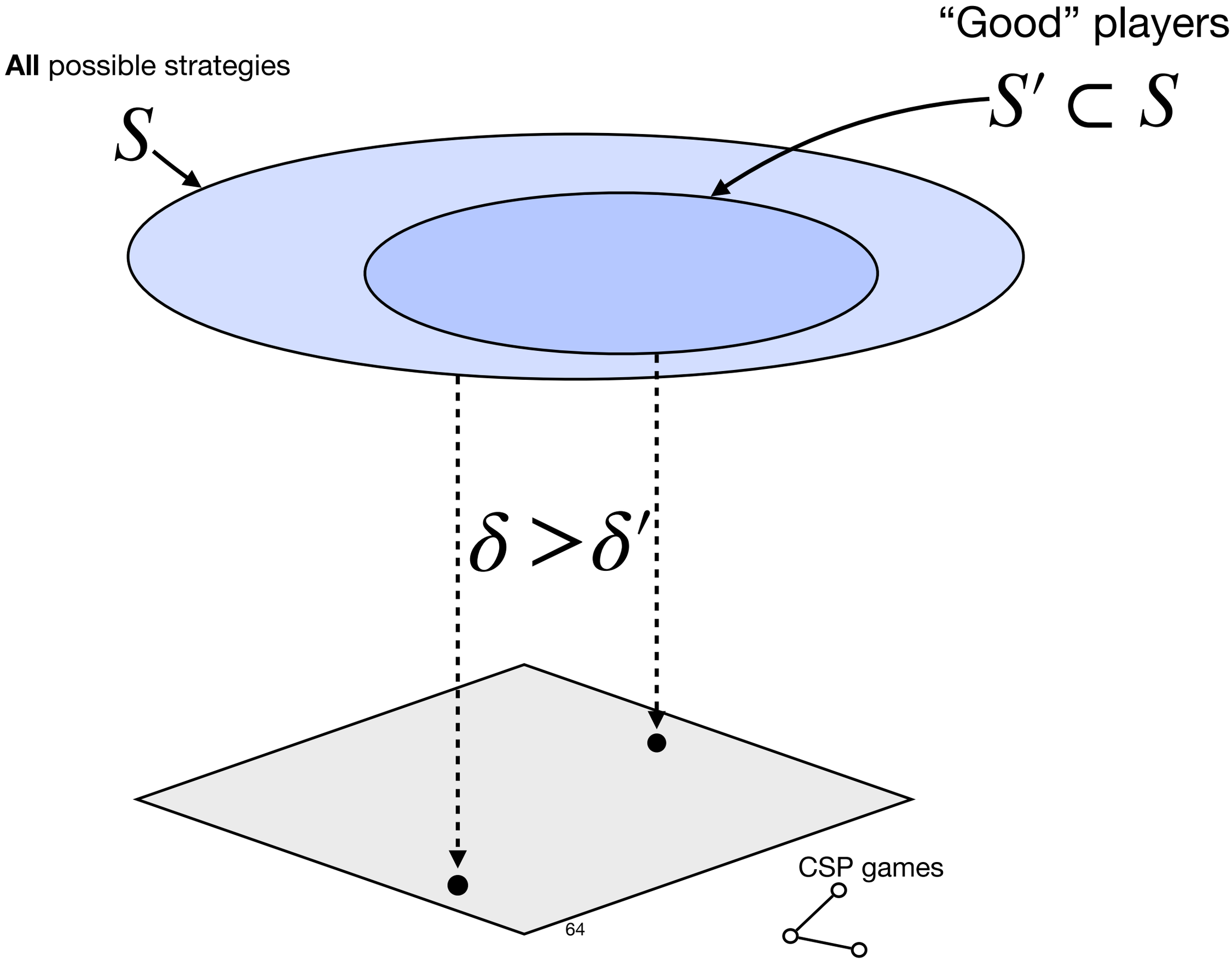


If they are well-approximated by a subgame stable CSP game!



Experiments

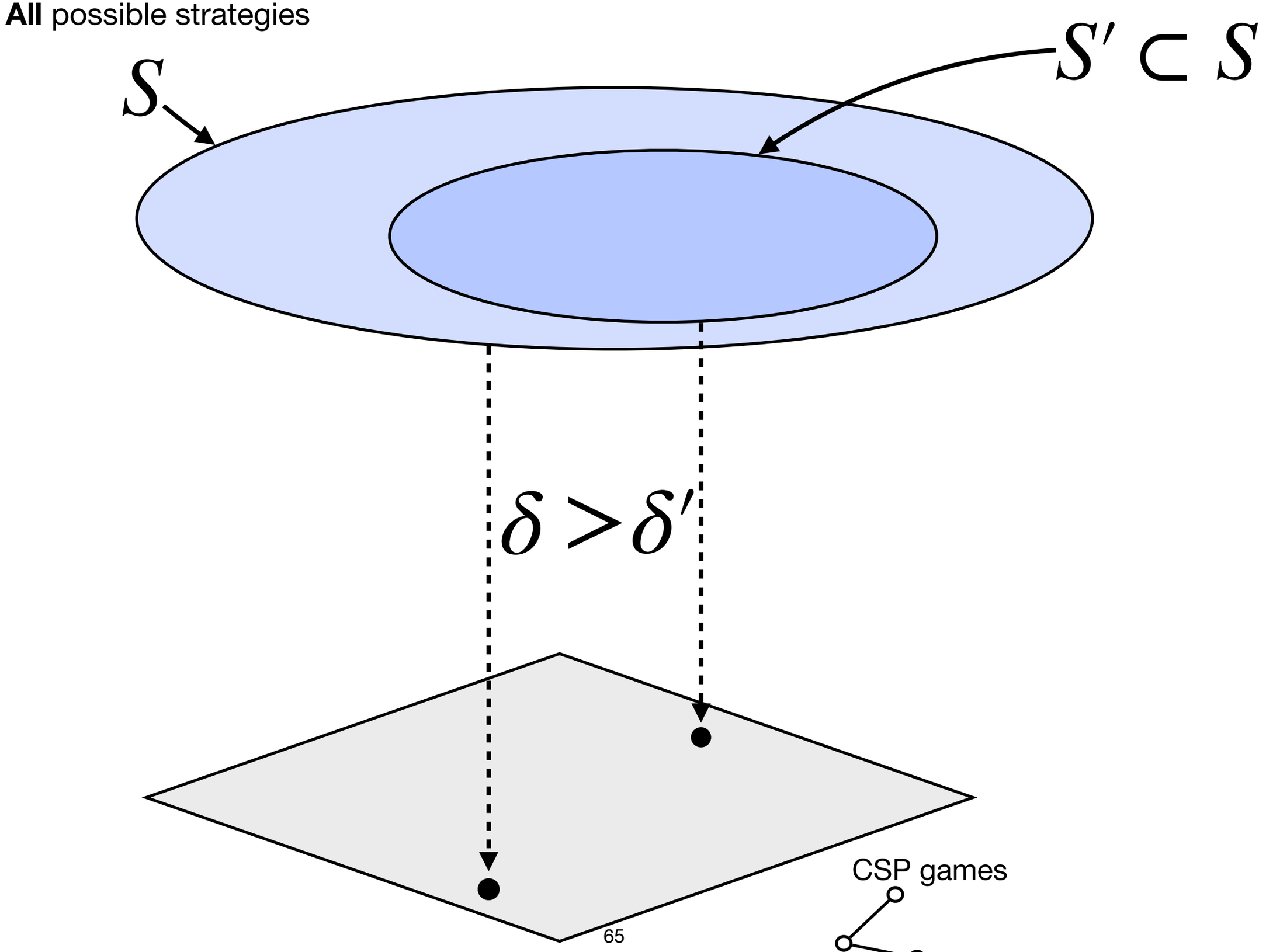
Conjecture. No limit Texas hold ‘em is approximately **constant-sum polymatrix** and **subgame stable**.


“Good” players
$S' \subset S$
All possible strategies
S
$\delta > \delta'$
CSP games

Learnable in **self-play**

~~"Good" players~~

All possible strategies



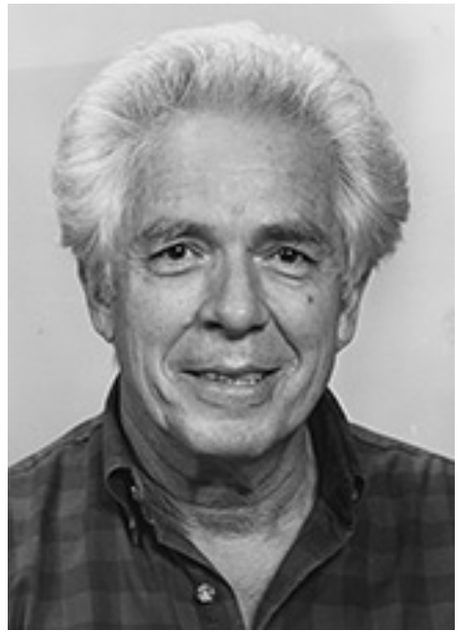
Conjecture. No limit Texas hold ‘em is approximately **constant-sum polymatrix** and **subgame stable**.

Conjecture. No limit Texas hold ‘em is approximately **constant-sum polymatrix** and **subgame stable…**

in parts of the game space that are actually played.

Kuhn Poker

Kuhn, 1950; Abou Risk & Szafron, 2010



- 3 players
- 4 cards, 1 round of betting with fixed bet size
- CFR was previously shown to converge to Nash equilibria in Kuhn poker

Hypothesis. **Kuhn poker** is approximately **constant-sum polymatrix** and **subgame stable…**

in parts of the game space that are learnable by self-play.

Experiment Design

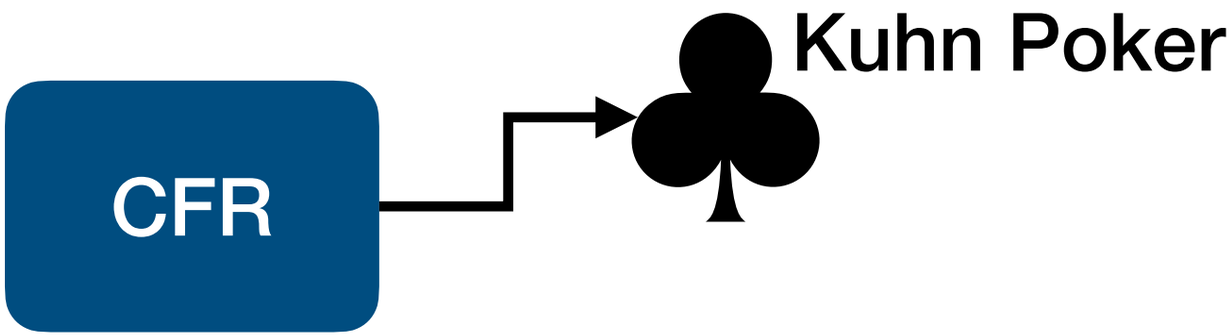
- Use a self-play algorithm to generate a set of strategies for each player

- Compute a set of match-ups between all these strategies.

- Check if this set of strategy profiles is approximately subgame-stable and constant-sum polymatrix.

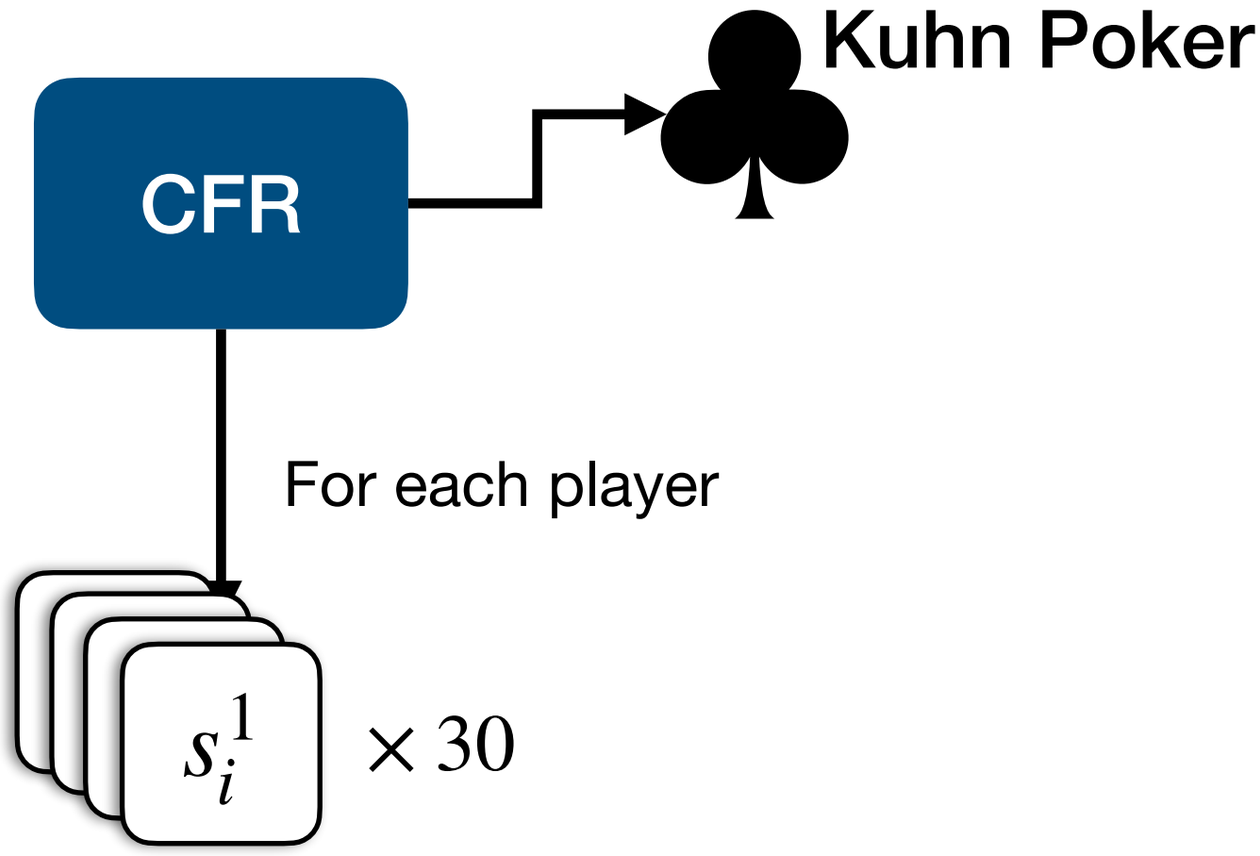
Hypothesis. **Kuhn poker** is approximately **constant-sum polymatrix** and **subgame stable…**

in parts of the game space that are learnable by self-play.



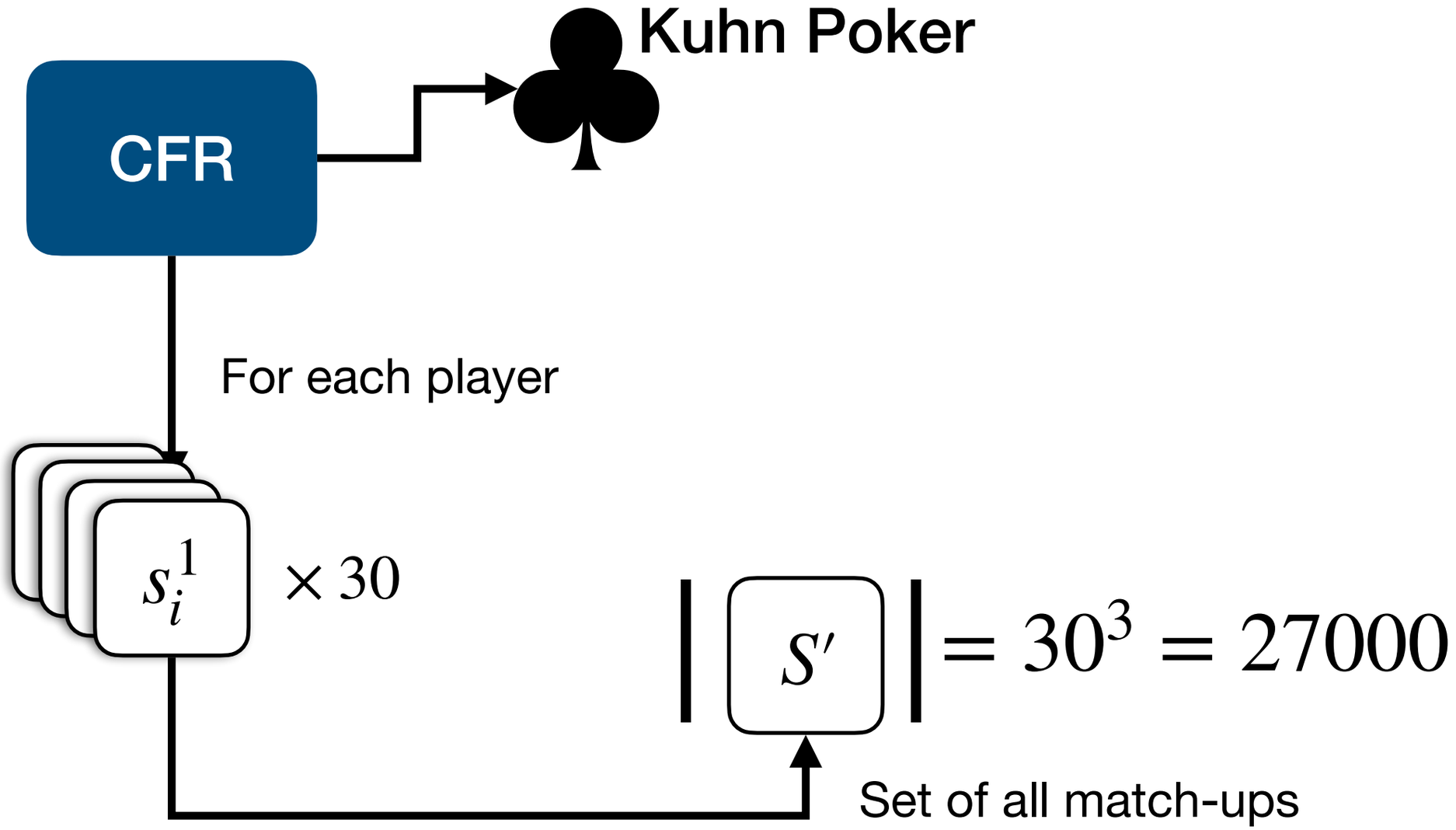
Hypothesis. **Kuhn poker** is approximately **constant-sum polymatrix** and **subgame stable…**

in parts of the game space that are learnable by self-play.



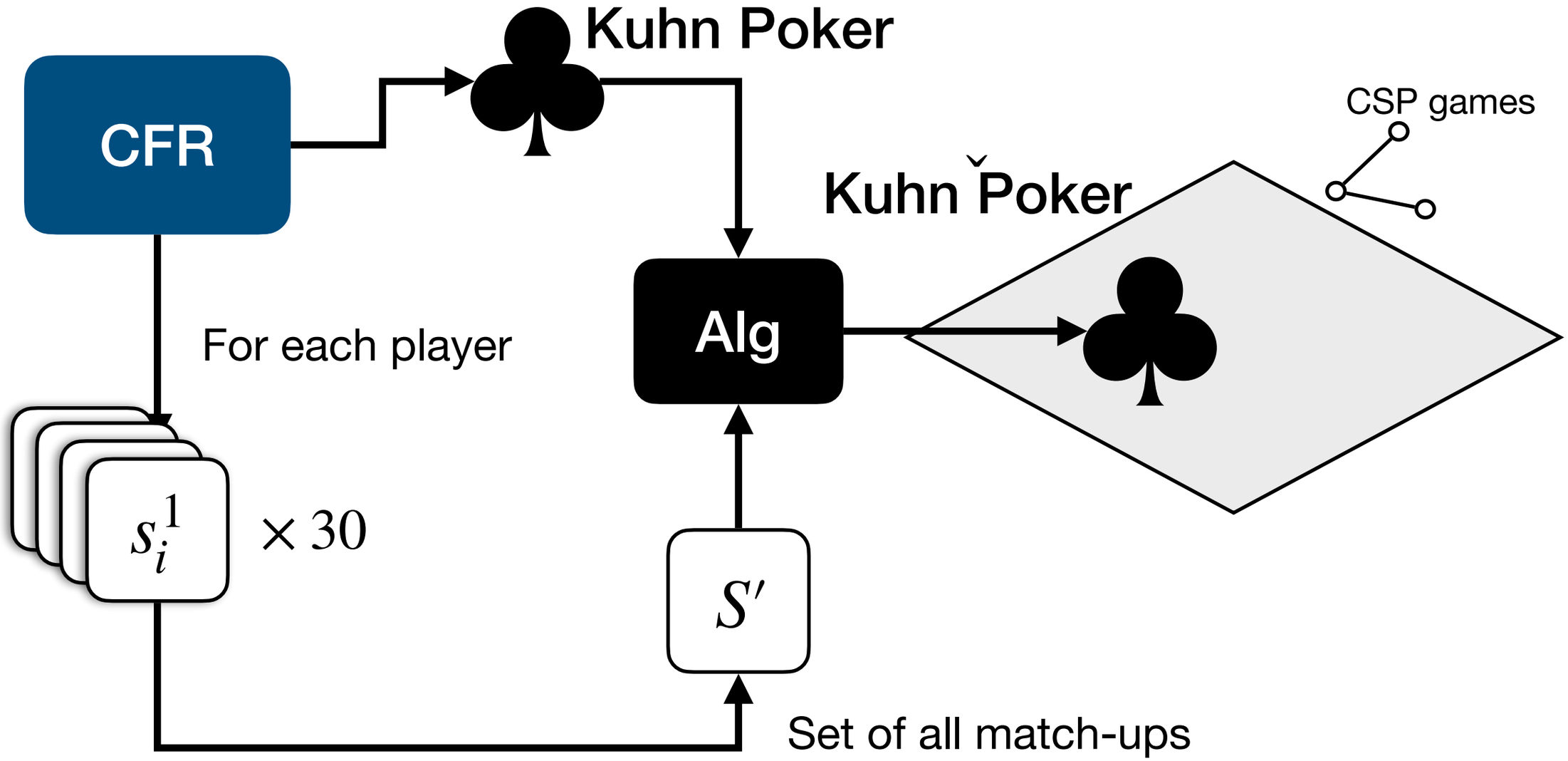
Hypothesis. **Kuhn poker** is approximately **constant-sum polymatrix** and **subgame stable…**

in parts of the game space that are learnable by self-play.



Hypothesis. **Kuhn poker** is approximately **constant-sum polymatrix** and **subgame stable…**

in parts of the game space that are learnable by self-play.



Results

Hypothesis. **Kuhn poker** is approximately **constant-sum polymatrix** and **subgame stable…**

in parts of the game space that are learnable by self-play.

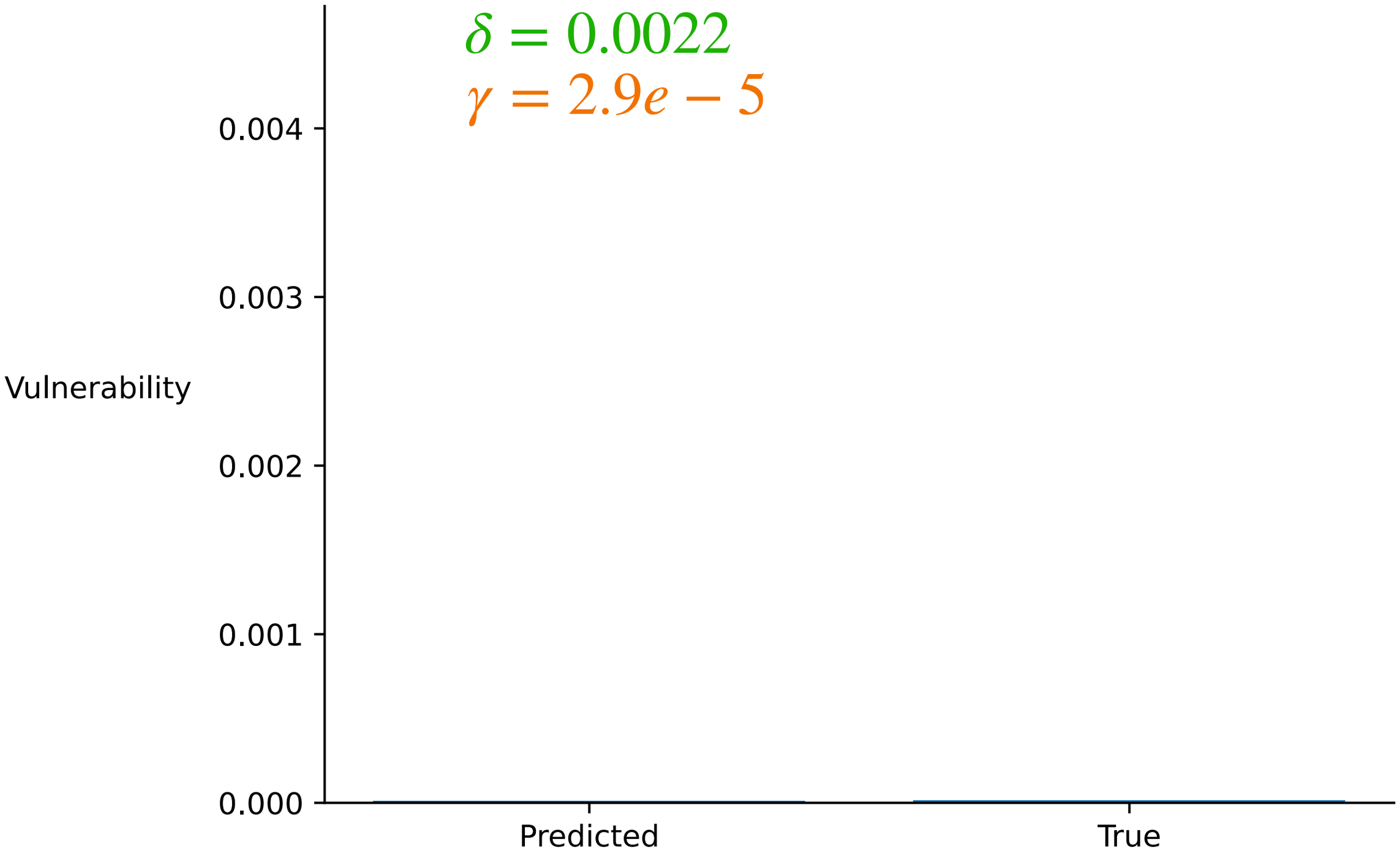
- $\max_{s \in S'} \left| u_i(s) - \check{u}_i(s) \right| \leq 0.0022$
- CFR strategies were $2.9e-5$-Nash of *each* subgame

Tightness of Bounds

Vulnerability $\leq 2(\delta + \gamma) = 0.0045$

$\delta = 0.0022$

$\gamma = 2.9e - 5$



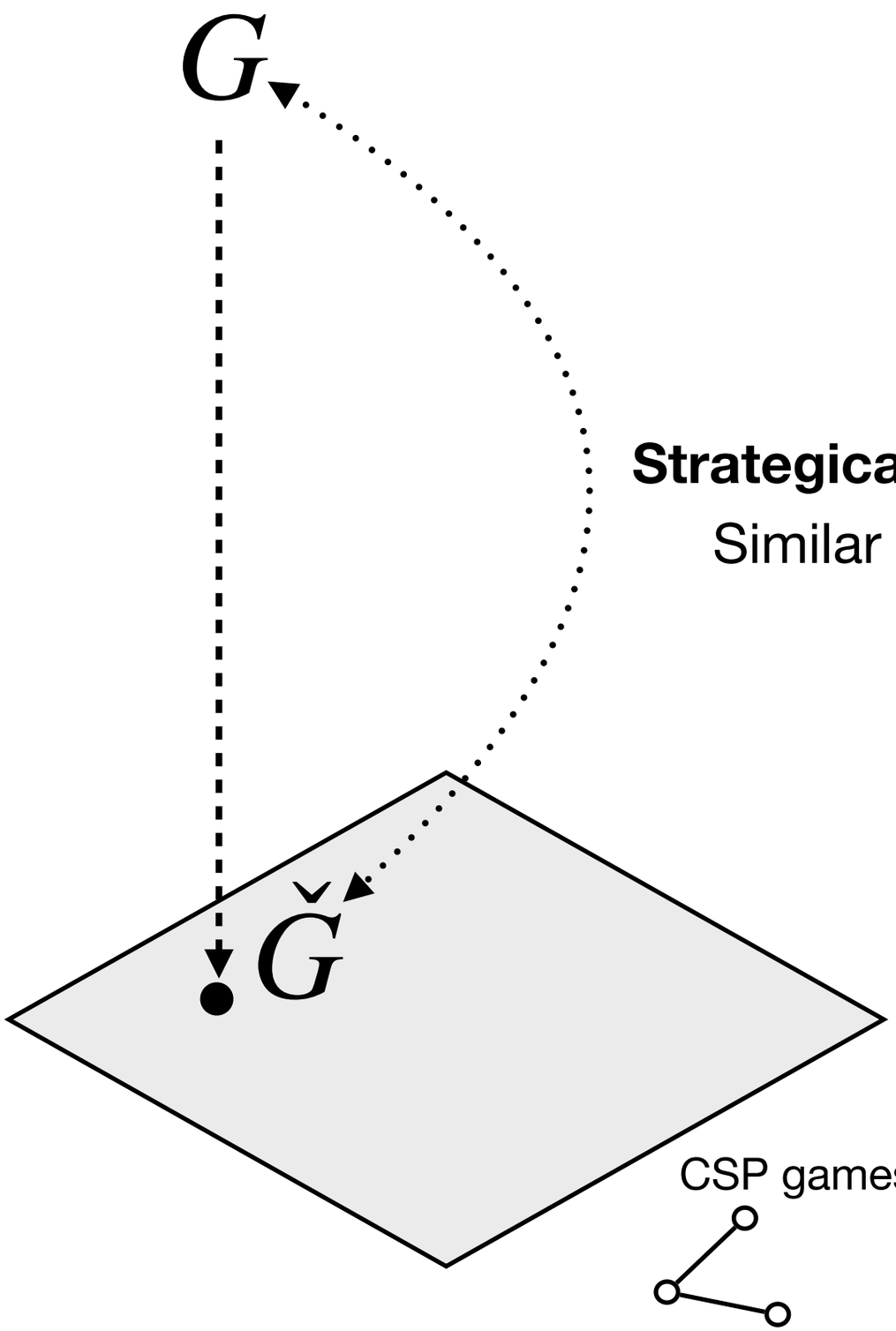
Other Experiments

Leduc Poker & Tiny Hanabi

- We repeated these experiments on **Leduc poker** and a **toy Hanabi** game.
- We had similar results to Kuhn poker for Leduc poker.
- We found that toy Hanabi was **not** well-approximated by a subgame stable CSP game

Generalizations

Strategic Equivalence



Definition. Strategic Equivalence
(Moulin & Vial, 1978)

For any s_{-i},

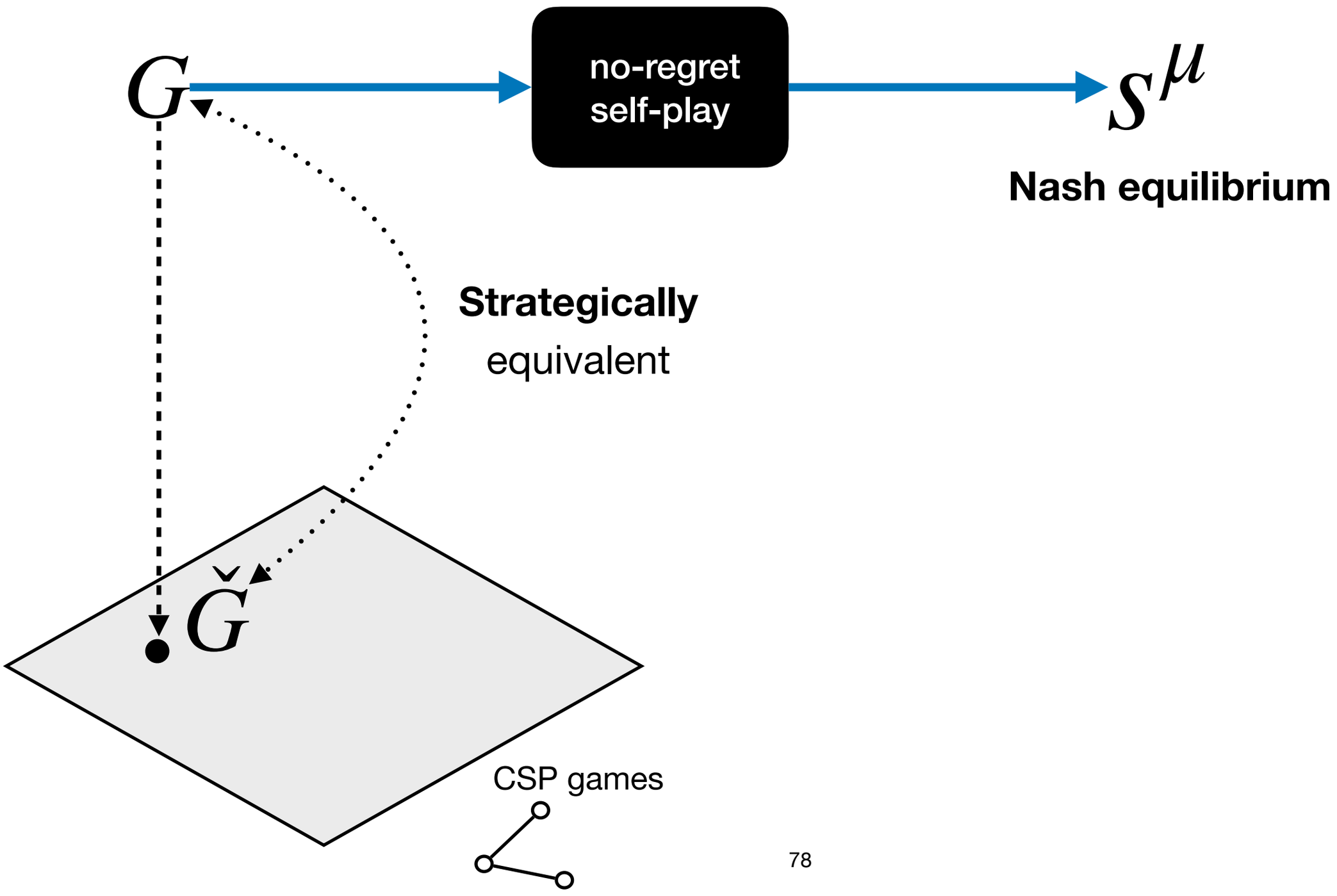
s_i' is preferred to s_i in G

$\Longleftrightarrow$

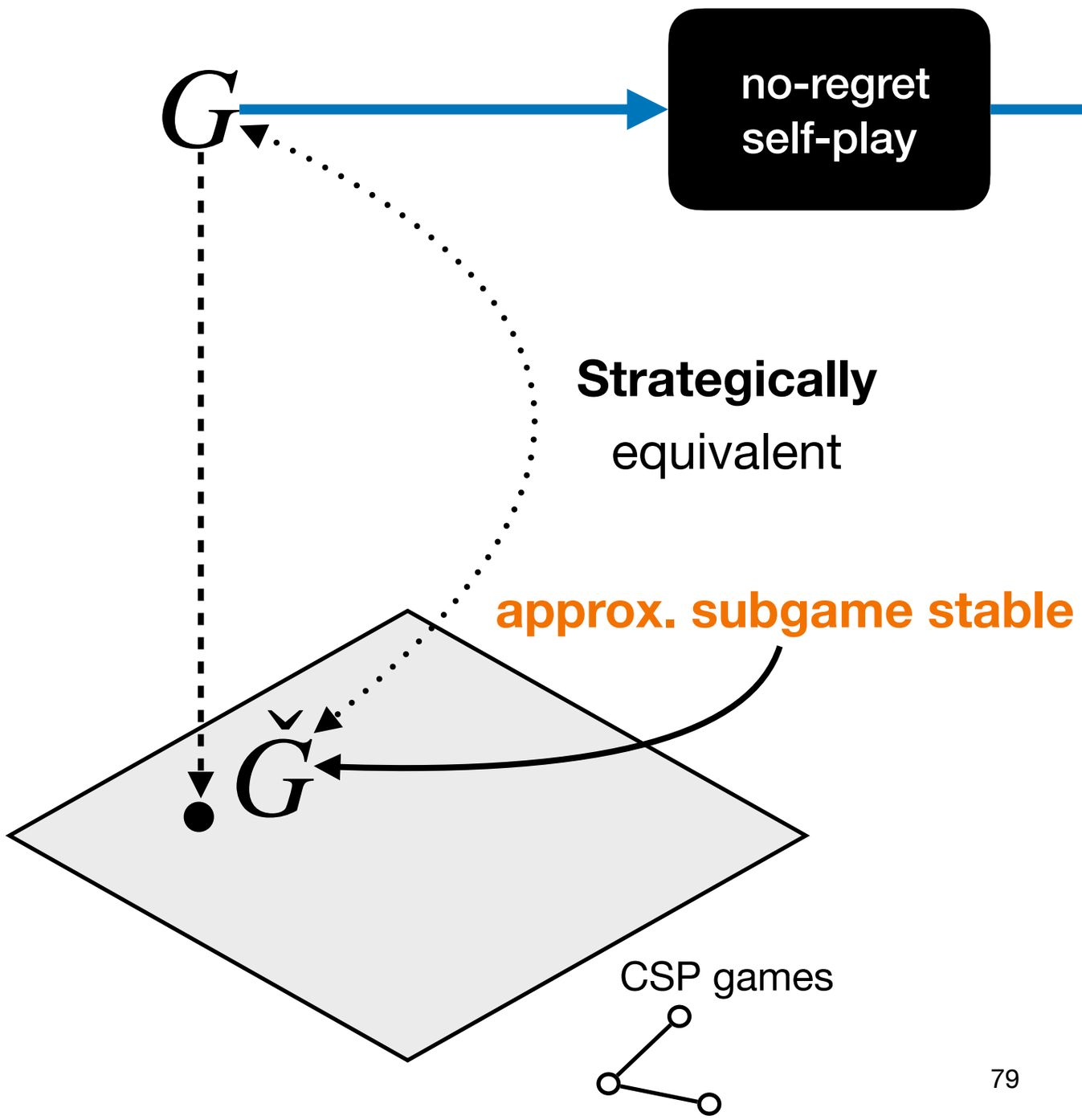
s_i' is preferred to s_i in $\check{G}$

Strategic Equivalence to CSP Games

- We generalize the algebraic characterization of **strategic equivalence** to **multiplayer games**



Strategic Equivalence to CSP Games



Self-Play Wishlist

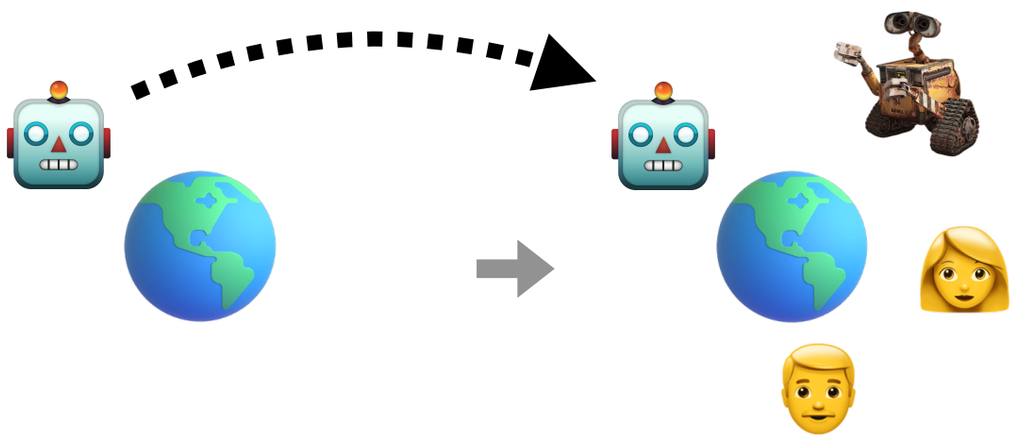
- ☐ ~~Low **vulnerability**~~
- ☐ ~~Similar **values**~~
- ☑ Nearly **Exchangeable**

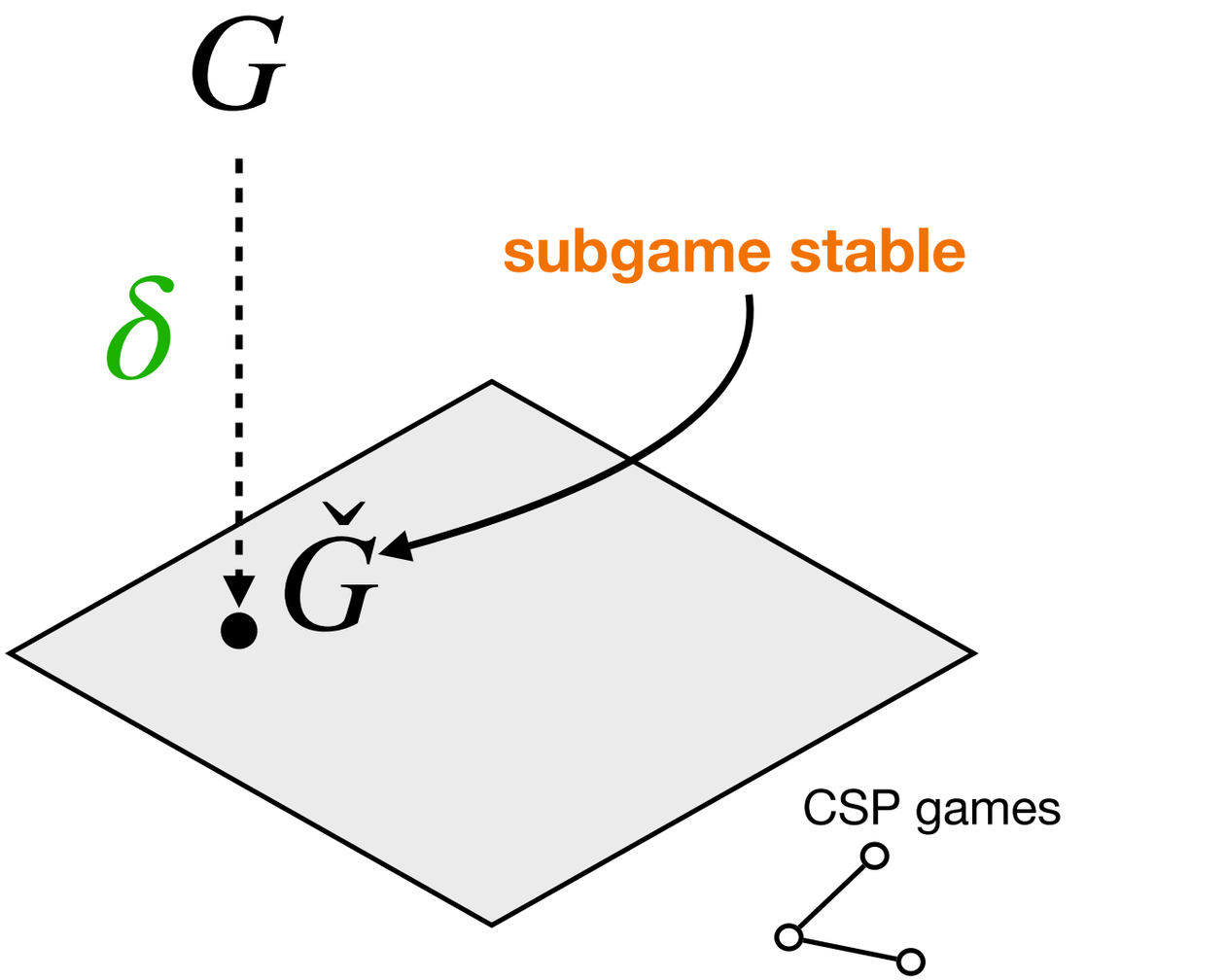
Strategic equivalence extends some (but not all) guarantees.

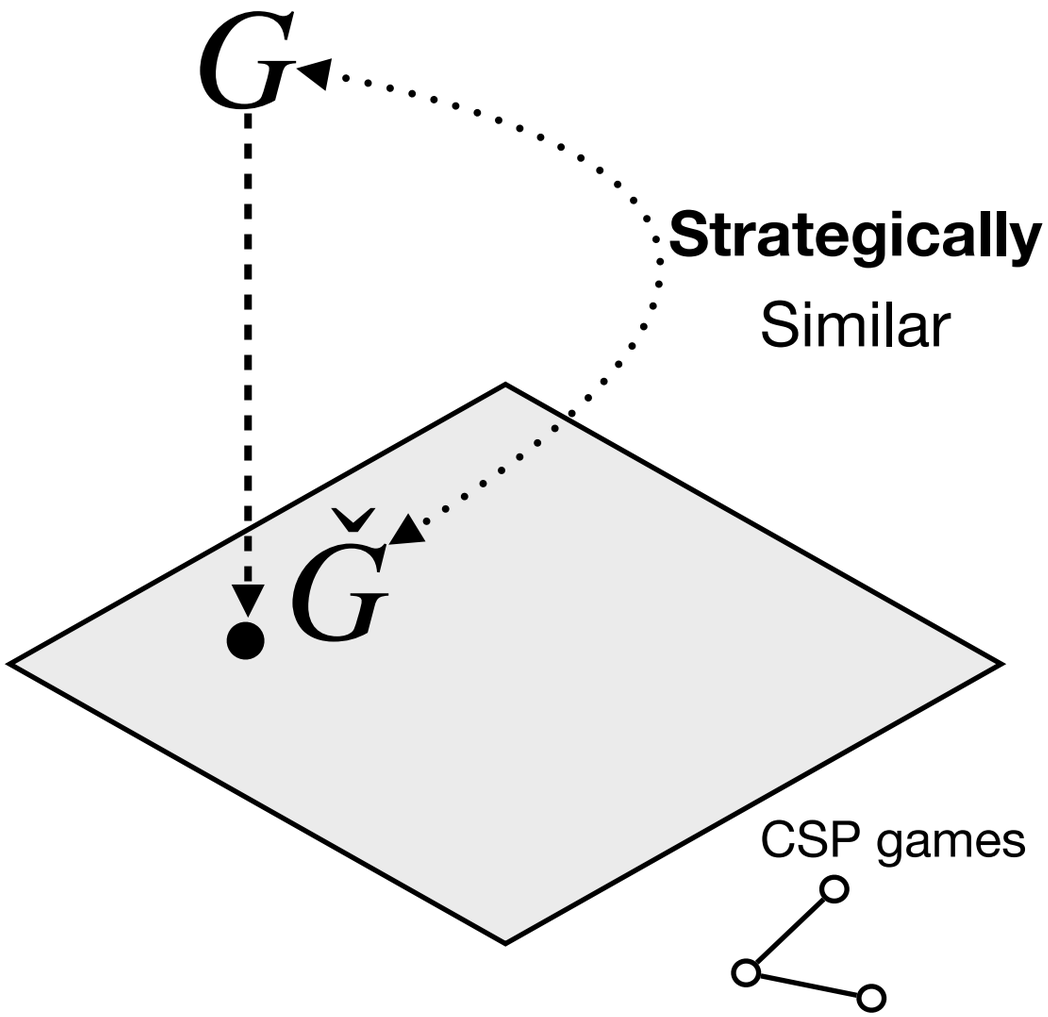
Strategically subgame stable CSP games generalize in a meaningful way strategically constant-sum games from Moulin & Vial.

Conclusion

Conclusion







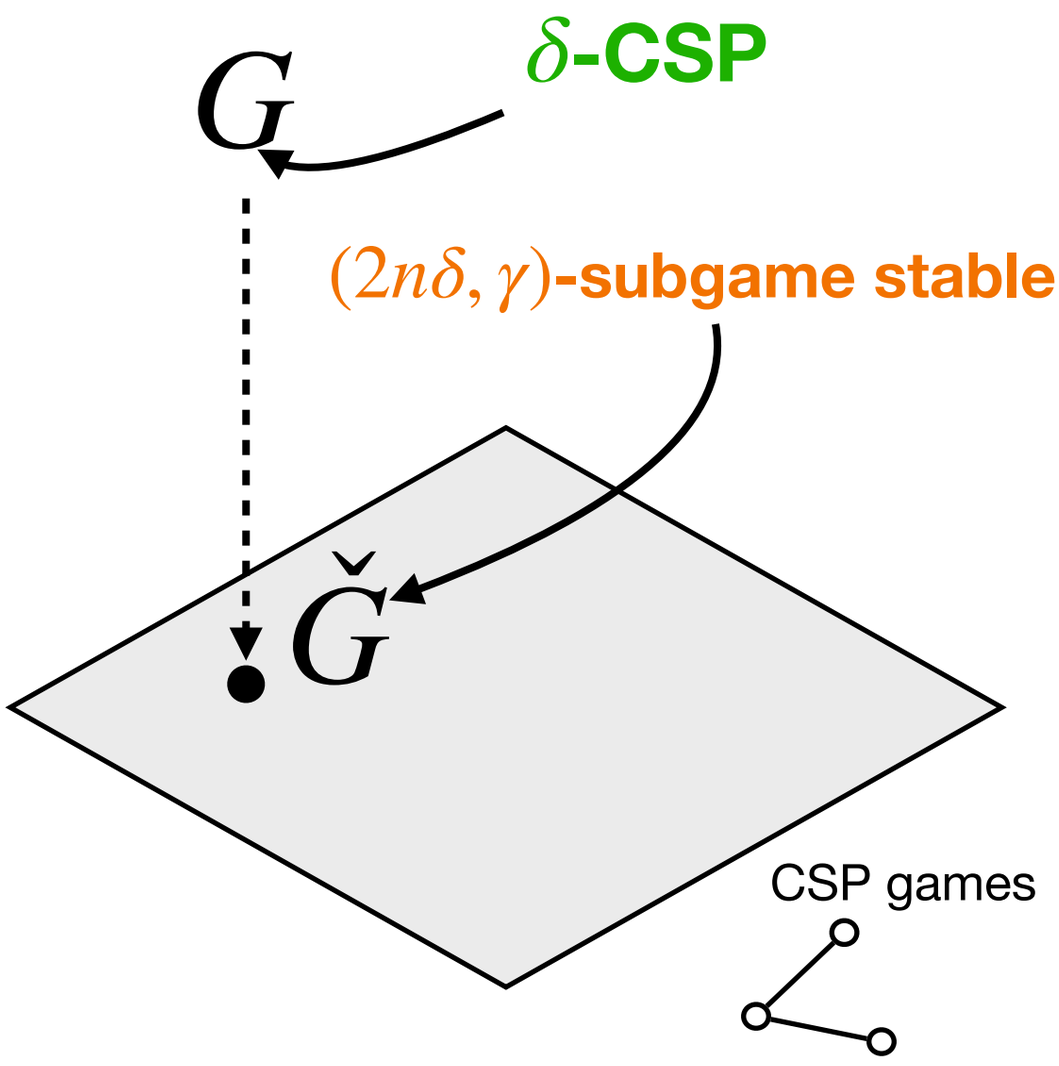
Self-Play Wishlist

- ☑ Low **vulnerability**
- ☑ Similar **values**
- ☑ Nearly **Exchangeable**

Appendix

Main Result

Guarantees for Multi-player Games



Theorem. Given G, if there exists a CSP game $\check{G}$ such that:

1. $\max_{s \in S} \left| u_i(s) - \check{u}_i(s) \right| \leq \delta$

2. $\check{G}$ is $(2n\delta, \gamma)$-subgame stable

Then…

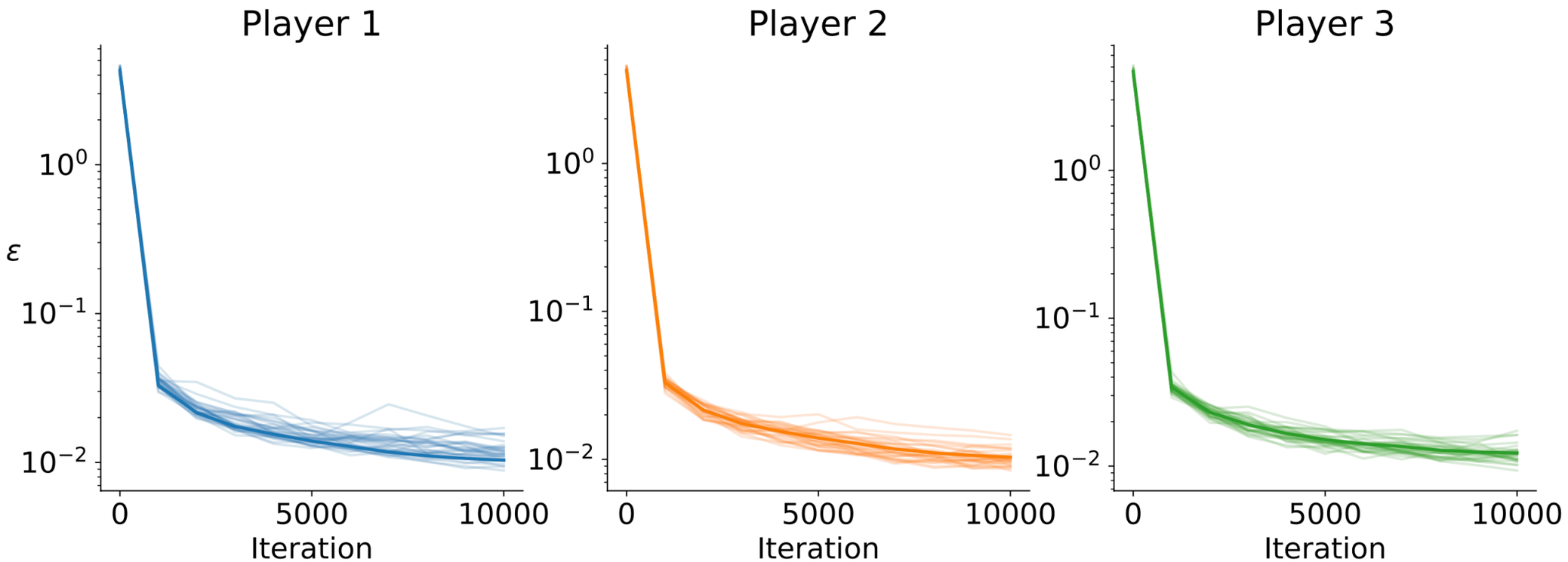
1. $\text{Vul}_i(s^\mu, S_{-i}) \leq |E_i|\gamma + 2\delta$

2. $\exists v_i : |v_i - u_i(s^\mu)| \leq |E_i|\gamma + \delta$

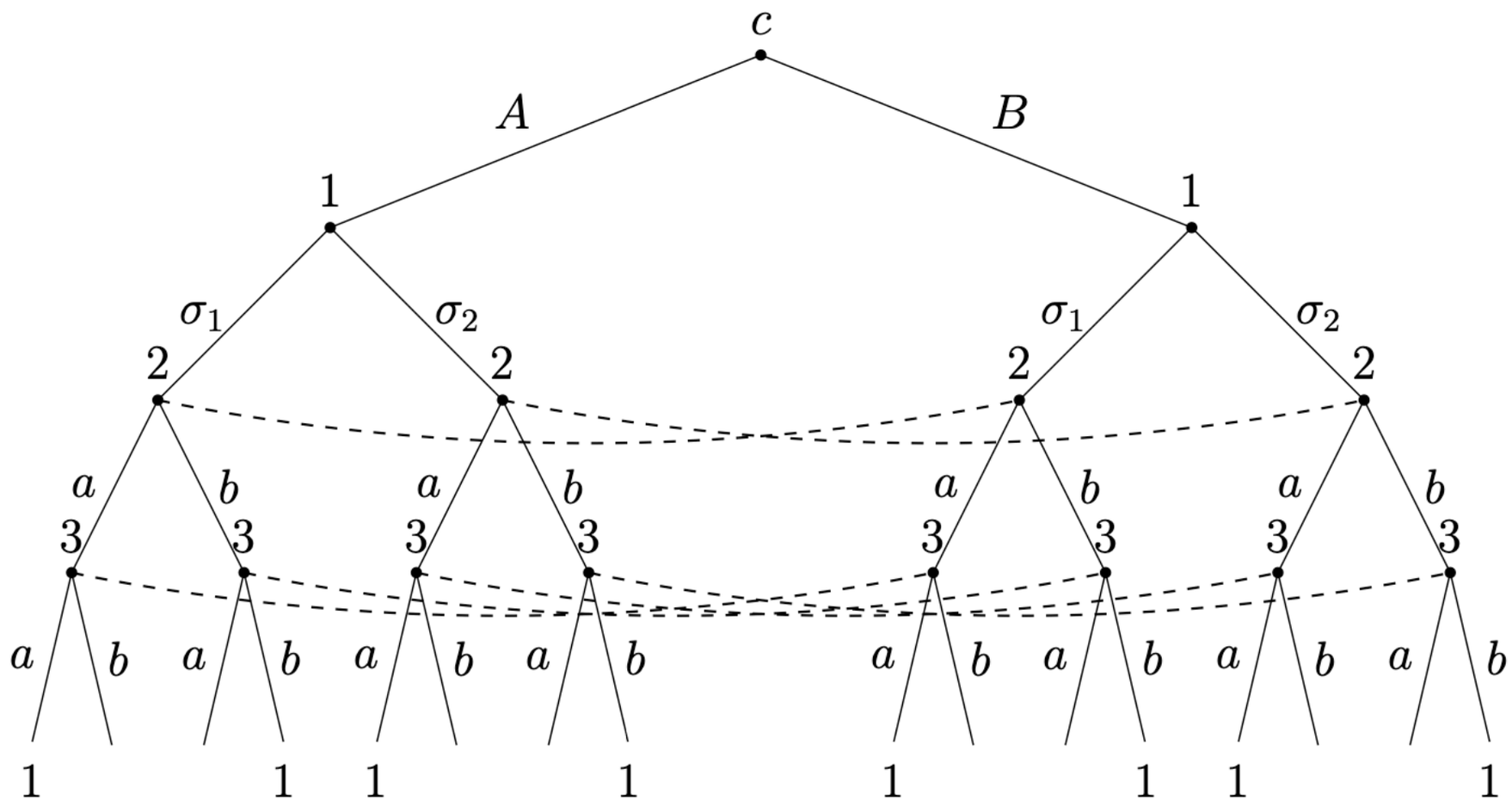
3. If s is a profile where each strategy is a marginal strategy from a *different* CCE…

   Then s is a $|E_i|\gamma + 2\delta$-Nash equilibrium

CFR Computes Approximate Nash in Leduc Poker



Tiny Hanabi


c
A
B
1
1
σ1
σ2
σ1
σ2
2
2
2
2
a
b
a
b
a
b
a
b
3
3
3
3
3
3
3
3
a
b
a
b
a
b
a
b
a
b
a
b
a
b
1
1
1
1
1
1
1
1

Strategic Equivalence

Definition 6.1.1 (Strategic Equivalence)**.** G is strategically equivalent (SE) to G' for player i if

$$u_i(s_i', s_{-i}) \geqslant u_i(s_i, s_{-i}) \iff u_i'(s_i', s_{-i}) \geqslant u_i'(s_i, s_{-i}) \quad \forall s_i, s_i' \in S_i, s_{-i} \in S_{-i}.$$

$$u_i(\rho) = \lambda_i u_i'(\rho) + \hat{u}_i(\rho_{-i}) \quad \forall \rho \in \mathrm{P},$$

Loss Functions (1)

$$\mathcal{L}^{\delta}\left(\pi; \check{u}, u\right) \doteq \sum_{i \in N} \left|\check{u}_i(\pi) - u_i(\pi)\right|$$

Loss Functions (2)

$$\mathcal{L}_{ij}^{\gamma}(\pi_{ij}, \pi_{ij}^{*}; \check{u}) \doteq \max\left(\check{u}_{ij}(\pi_i^{*}, \pi_j) - \check{u}_{ij}(\pi_{ij}), 0\right)$$
$$+ \max\left(\check{u}_{ji}(\pi_i, \pi_j^{*}) - \check{u}_{ji}(\pi_{ij}), 0\right) .$$

$$\mathcal{L}^{\gamma}\left(\pi, \pi^{*}; \check{u}\right) \doteq \sum_{(i,j)\in E} \mathcal{L}_{ij}^{\gamma}(\pi_{ij}, \pi_{ij}^{*}; \check{u})$$

Loss Functions (3)

$$\mathcal{L}(\Pi^{\delta}, \Pi^{\gamma}, \Pi^{*}; \breve{u}, u) \doteq \frac{\lambda}{B_1} \sum_{\pi \in \Pi^{\delta}} \mathcal{L}^{\delta}(\pi; \breve{u}, u) + \frac{(1-\lambda)}{B_2} \sum_{\pi \in \Pi^{\gamma}} \sum_{\pi^{*} \in \Pi^{*}} \mathcal{L}^{\gamma}(\pi, \pi^{*}; \breve{u})$$

$$\breve{u} \leftarrow \breve{u} - \eta \cdot \nabla_{\breve{u}} \mathcal{L}(\Pi^{\delta}, \Pi^{\gamma}, \Pi^{*}; \breve{u}, u).$$

Sample-Based Algorithm

Algorithm 2 Compute $\check{G}$

Input: G, Π', η, T_{in}, T_{out}, λ, B_1, B_2
Initialize $\check{u}$ to all 0
$\Pi^{\times} \leftarrow \times_{i \in N} \hat{\Pi}_i$
for $t_{out} \in 1...T_{out}$ **do**
  $\Pi^{*} \leftarrow \text{getBRs}(\check{G}, \Pi')$
  for $t_{in} \in 1...T_{in}$ **do**
    $\Pi^{\delta} \leftarrow$ sample batch of size B_1 u.a.r. from $\Pi^{\times}$
    $\Pi^{\gamma} \leftarrow$ sample batch of size B_2 u.a.r. from Π'
    $g \leftarrow \nabla_{\check{u}} \mathcal{L}(\Pi^{\delta}, \Pi^{\gamma}, \Pi^{*}; \check{u}, u)$
    $\check{u} \leftarrow \check{u} - \eta \cdot \frac{g}{\|g\|_2}$
  end for
end for
{Lastly, output δ and γ}
$\delta \leftarrow \max_{\pi \in \Pi^{\times}} |u_i(\pi) - \check{u}_i(\pi)|$
$\gamma \leftarrow \max_{\pi \in \Pi'} \max_{i \neq j \in N \times N} (\check{u}_{ij}(BR_{ij}(\pi_j), \pi_j) - \check{u}_{ij}(\pi_i, \pi_j))$
return $\check{u}, \gamma, \delta$

Compute Subgame Stability (1)

Algorithm 1 Compute γ

Input: $G = (N, E, \mathrm{P}, u)$, a polymatrix game
$\gamma \leftarrow -\infty$
for $(i, j) \in E$ **do**
  for $\rho'_i \in \mathrm{P}_i$ **do**
    if $\mathrm{LP1}(i, j, \rho'_i)$ not infeasible **then**
      $\gamma_{ij}^{\rho'_i} \leftarrow \mathrm{LP1}(i, j, \rho'_i)$
      $\gamma \leftarrow \max(\gamma, \gamma_{ij}^{\rho'_i})$
    end if
  end for
  for $\rho'_j \in \mathrm{P}_j$ **do**
    if $\mathrm{LP1}(j, i, \rho'_j)$ not infeasible **then**
      $\gamma_{ji}^{\rho'_j} \leftarrow \mathrm{LP1}(j, i, \rho'_j)$
      $\gamma \leftarrow \max(\gamma, \gamma_{ji}^{\rho'_j})$
    end if
  end for
end for
return γ

Computing CSP Decompositions

$$
\begin{aligned}
\min \quad & \underline{\delta} \\
\text{s.t.} \quad & u_i(\rho) - \sum_{j \in -i} \check{u}_{ij}(\rho_i, \rho_j) \leqslant \underline{\delta} \quad \forall i \in N, \rho \in \mathrm{P} \\
& u_i(\rho) - \sum_{j \in -i} \check{u}_{ij}(\rho_i, \rho_j) \geqslant -\underline{\delta} \quad \forall i \in N, \rho \in \mathrm{P} \\
& \check{u}_{ij}(\rho_i, \rho_j) + \check{u}_{ji}(\rho_i, \rho_j) = c_{ij} \quad \forall i \neq j \in N, (\rho_i, \rho_j) \in \mathrm{P}_{ij},
\end{aligned}
$$