# Prosocial or Selfish?

## Agents with different behaviors for Contract Negotiation using Reinforcement Learning

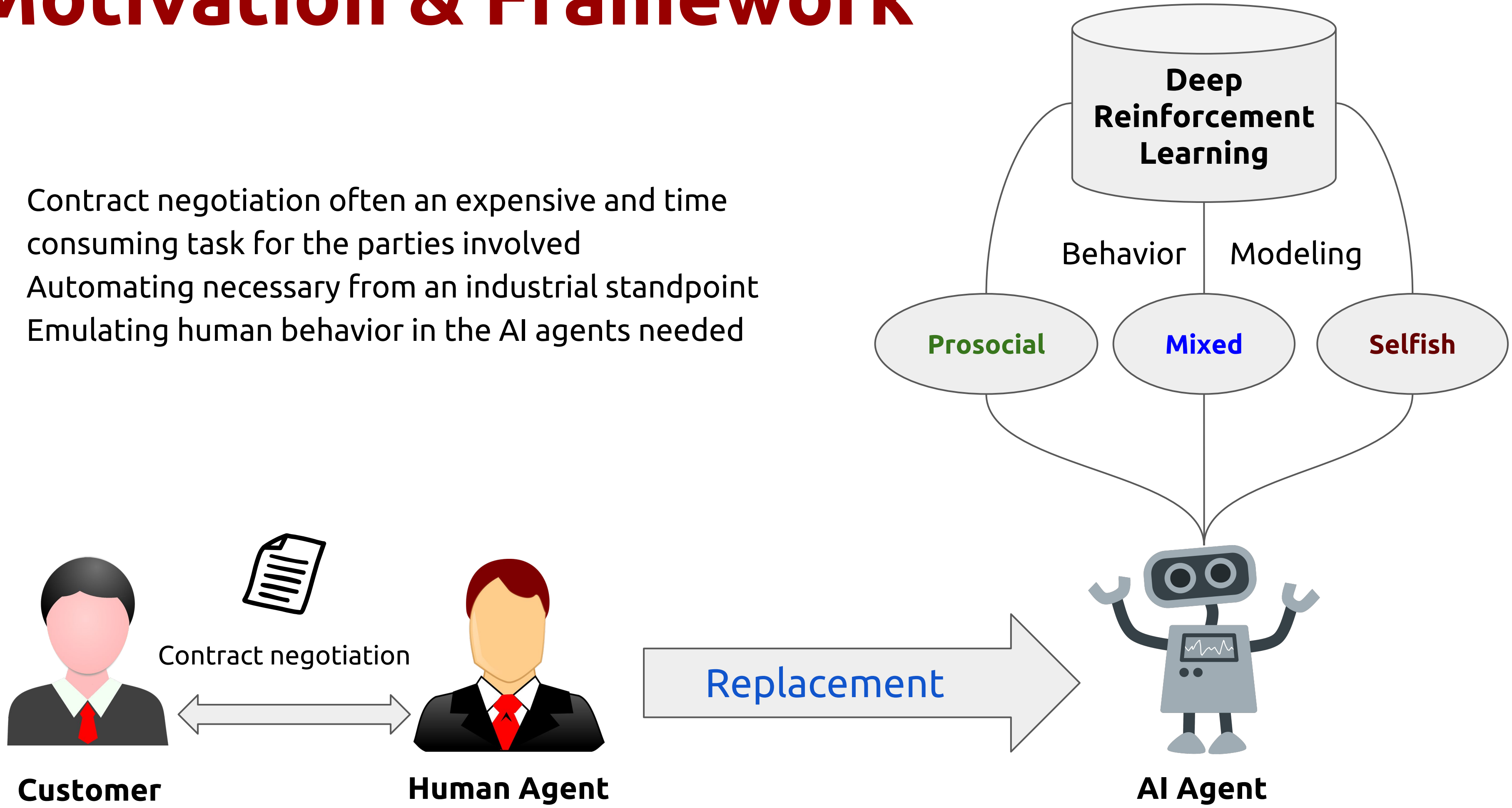
**Vishal Sunder**

Co-authors - Lovekesh Vig, Arnab Chatterjee, Gautam Shroff
**TCS Research, New Delhi, India**

**ACAN 2018**

# Motivation & Framework

- Contract negotiation often an expensive and time consuming task for the parties involved
- Automating necessary from an industrial standpoint
- Emulating human behavior in the AI agents needed



Deep Reinforcement Learning

Behavior Modeling

Prosocial

Mixed

Selfish

Contract negotiation

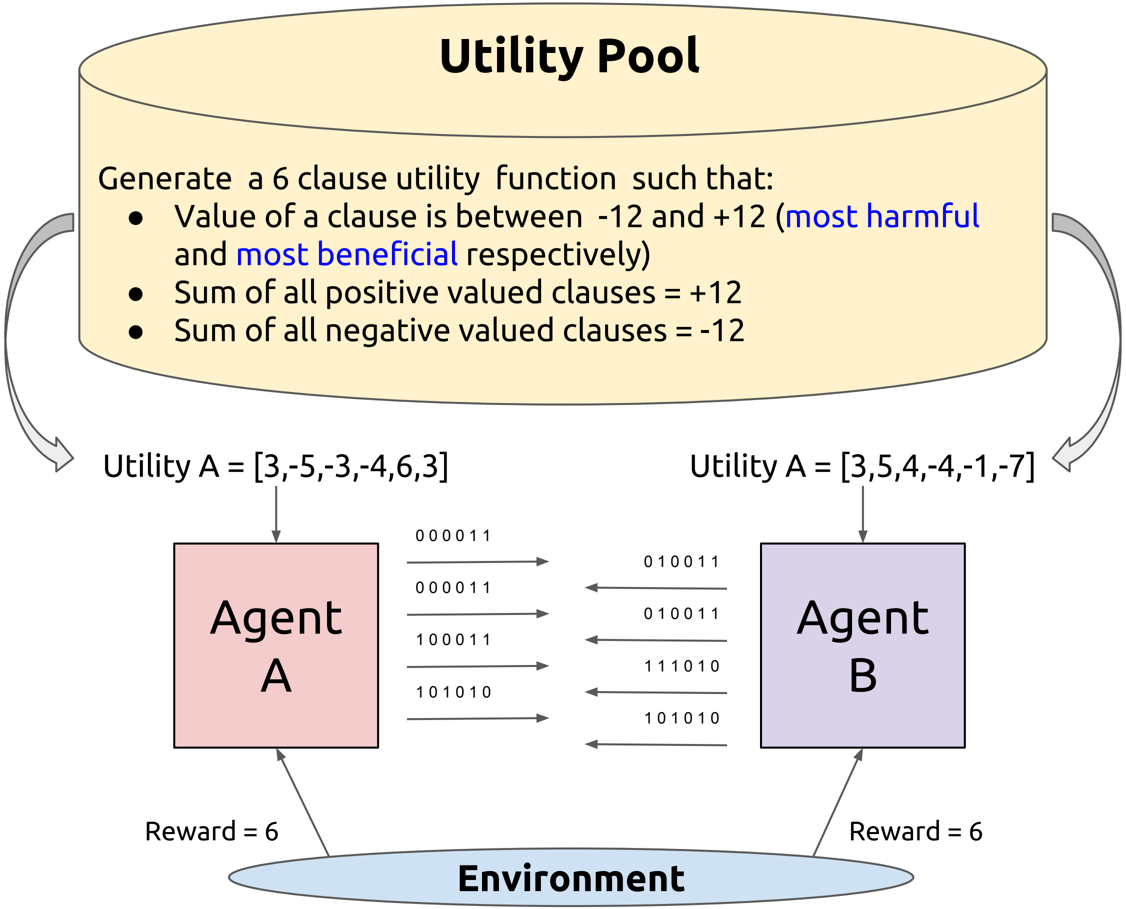
Replacement

Customer

Human Agent

AI Agent

# Negotiation Environment



**Utility Pool**

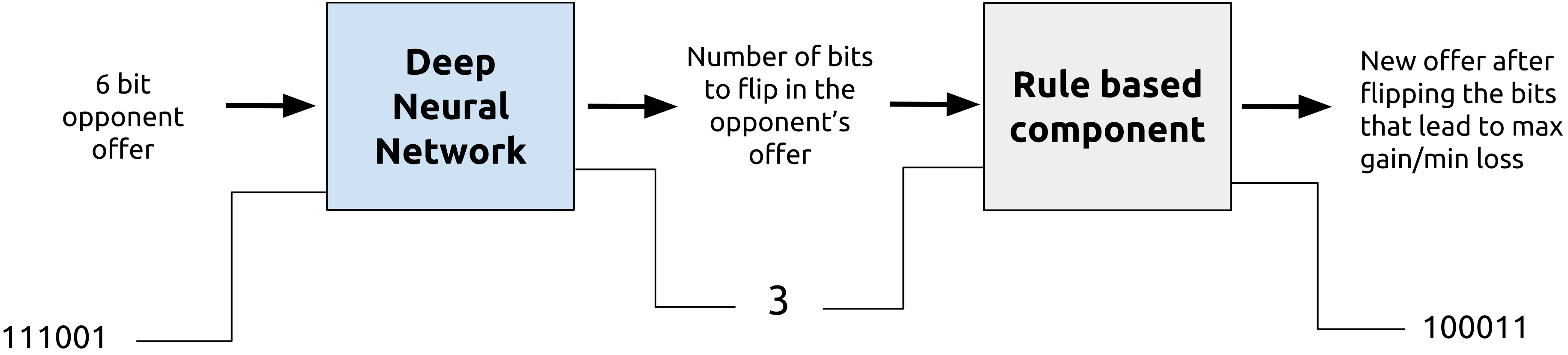
Generate a 6 clause utility function such that:

- Value of a clause is between -12 and +12 (most harmful and most beneficial respectively)
- Sum of all positive valued clauses = +12
- Sum of all negative valued clauses = -12

- Utility sampled uniformly from the pool.
- Communication follows a sequential structure.
- Communication goes on till:
  - An agent gives the same offer that it receives
  - Offer limit of 30 is reached

# Agent Modeling

## Two part model



6 bit opponent offer

Deep Neural Network

Number of bits to flip in the opponent's offer

Rule based component

New offer after flipping the bits that lead to max gain/min loss

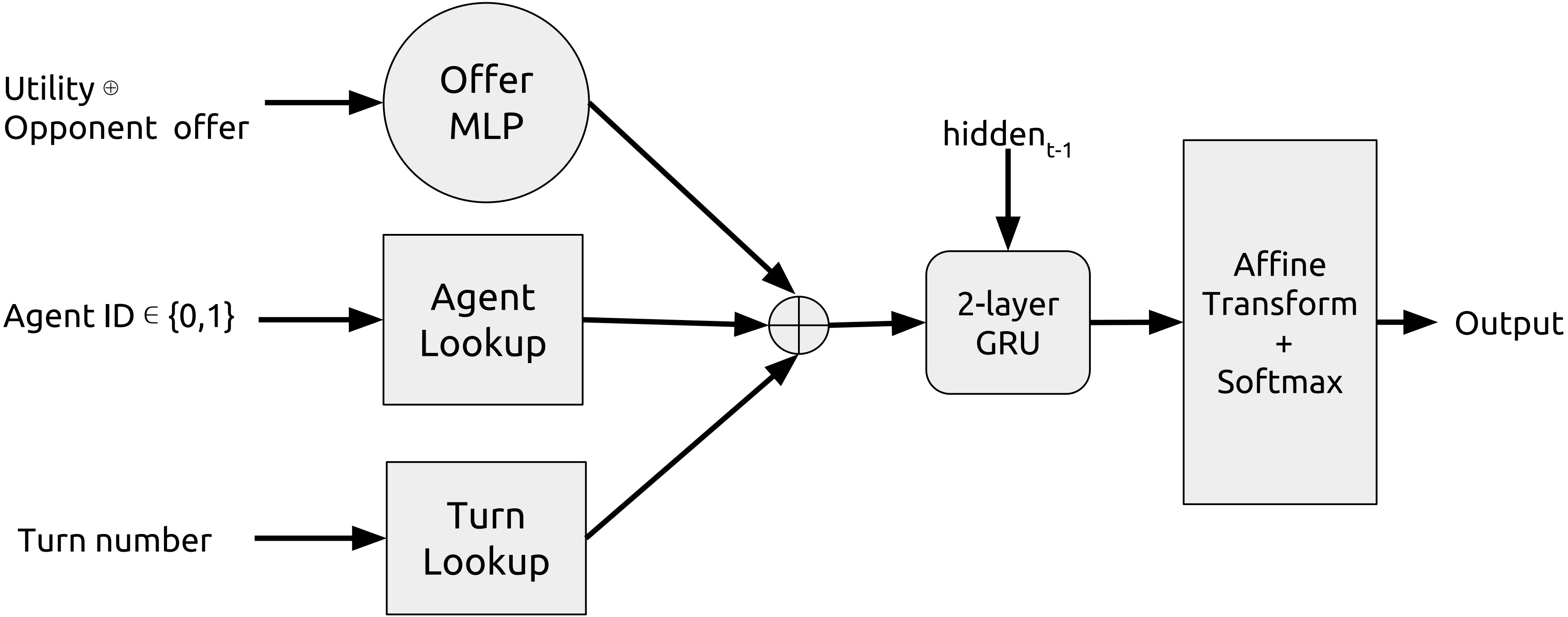
111001

3

100011

Utility = [2,-6,-2,-4,7,3]

# Deep Neural Network Model



# Training

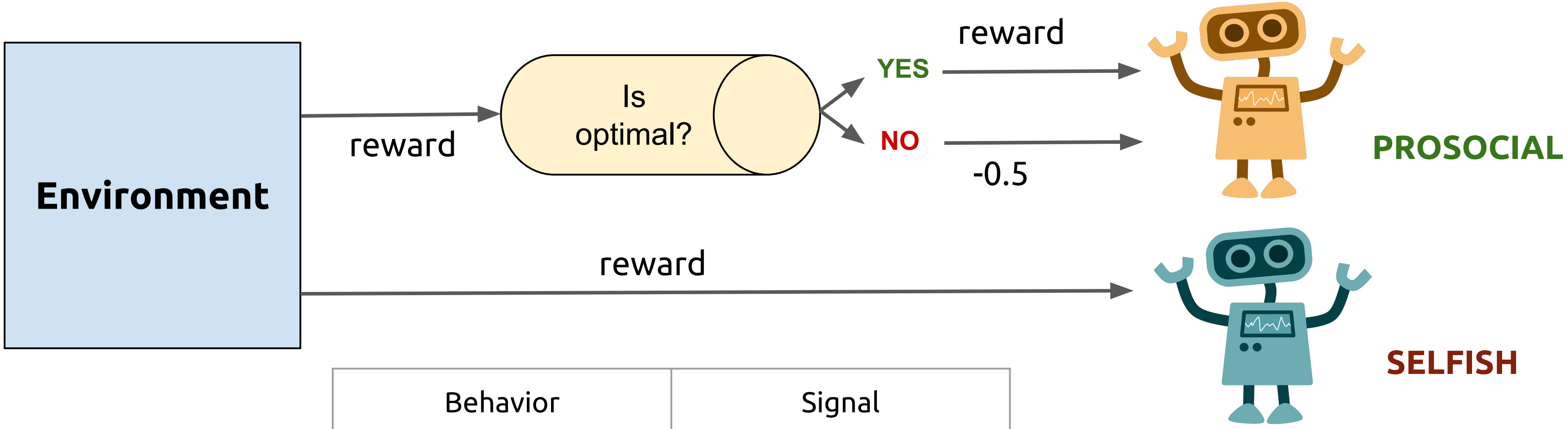
- After each episode (a negotiation game), each agent *i* ∈ {A,B} tries to maximize the following objective individually:

$$L_i = \mathop{\mathbb{E}}_{x_t \sim (\pi_A, \pi_B)} \left[ \sum_t \gamma^{(T-t)} (r_i(x_{1\ldots T}) - b_i) \right] + \lambda H[\pi_i].$$

- The deep model was trained by SGD with nesterov and momentum
- The gradient of $L_i$ is computed as in REINFORCE[1]

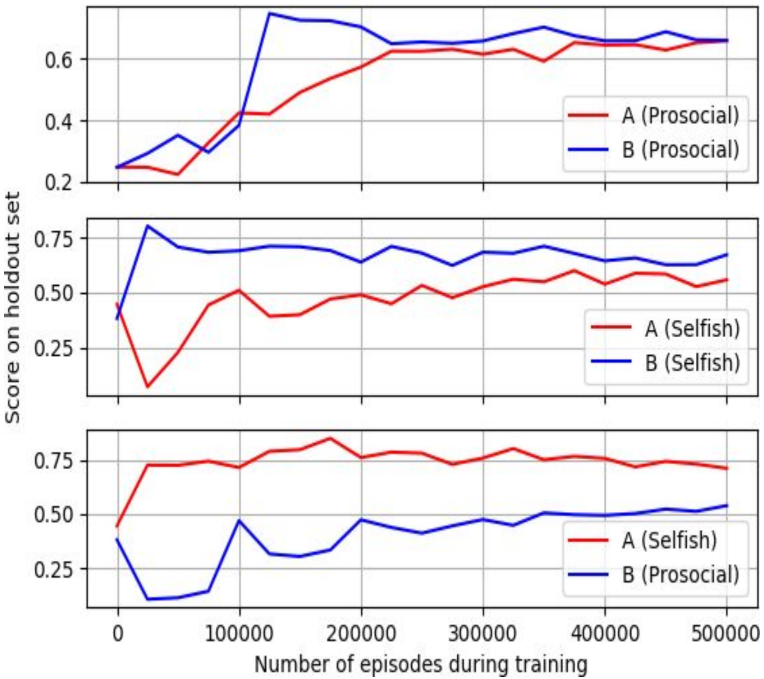
[1] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In Reinforcement Learning. Springer, 5–32.

# Modeling behavior through rewards



| Behavior | | Signal | |
|---|---|---|---|
| A | B | A | B |
| Prosocial (PP) | Prosocial (PP) | Yes | Yes |
| Selfish (SS) | Selfish (SS) | No | No |
| Prosocial (PS) | Selfish (SP) | Yes | No |

# Coordination between trained agents (I)



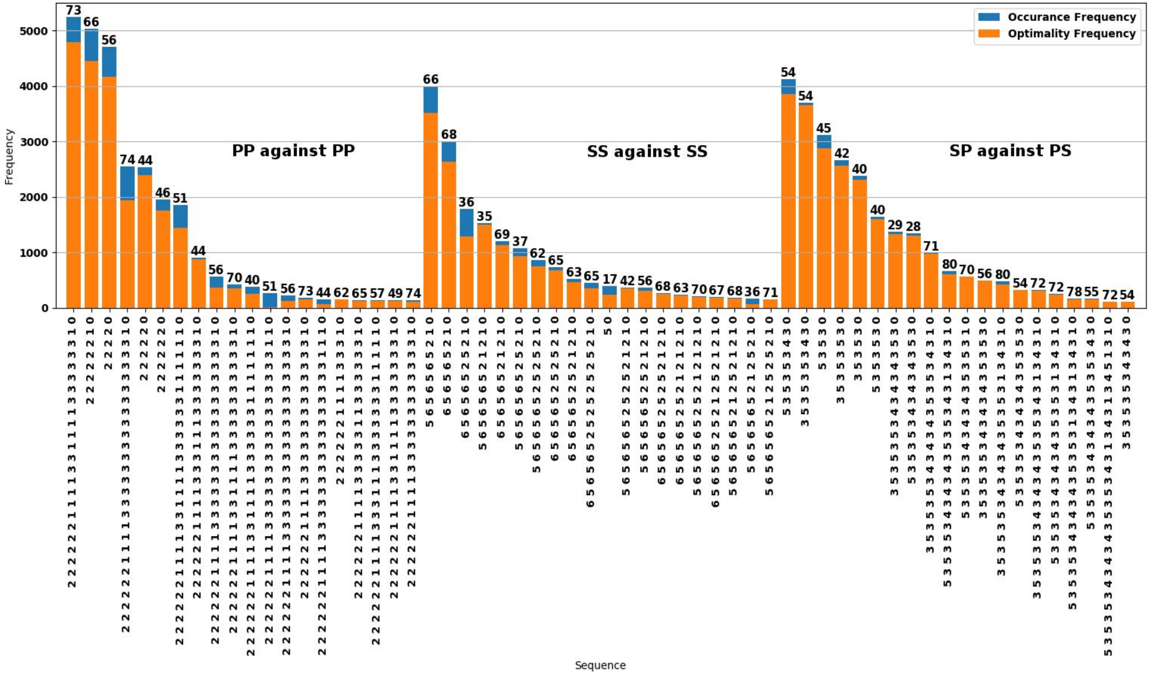
- Prosocial vs Prosocial:
    - Gradually converge to same score
    - Reach “middle ground”
- Selfish v Selfish:
    - One agent scores more than the other
    - One agent compromises in order to reach a deal
- Selfish v Prosocial:
    - Selfish outscores prosocial

# Coordination between trained agents (II)

| **Agent A** | **Agent B** | **Dialog Length** | **Agreement Rate (%)** | **Optimality Rate (%)** | **Average Scores** | |
|---|---|---|---|---|---|---|
| | | | | | **A** | **B** |
| ***BASELINE RANDOM*** | | *15.90* | *100* | *24.55* | *0.25* | *0.25* |
| ***BASELINE COMMON*** | | *3.77* | *79.54* | *70.39 (88.49)* | *0.50* | *0.50* |
| **PP** | **PP** | 16.98 | 96.24 | 82.33 (85.55) | 0.65 | 0.66 |
| **SS** | **SS** | 17.47 | 88.31 | 74.88 (84.79) | 0.54 | 0.69 |
| **SP** | **PS** | 13.87 | 91.90 | 86.74 (94.38) | 0.73 | 0.55 |

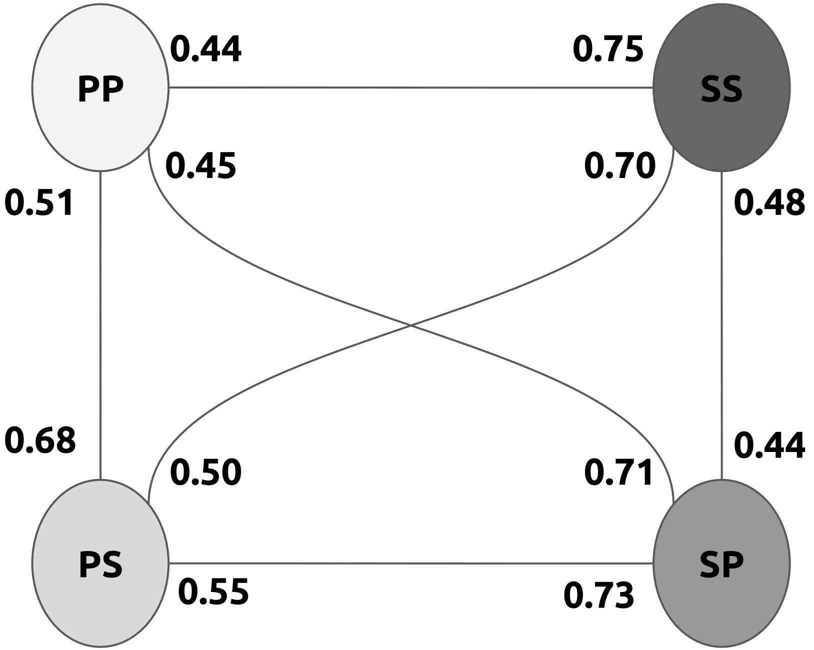
- All behavioral combinations do better than the baselines.
- Agents trained against each other are able to “coordinate” their moves.
- Joint reward maximum when both agents are prosocial.

# Coordination between trained agents (III)



- Joint preference for certain sequences more than others.
- Do the agents learn to identify context?:
  - The number on top of each bar is the optimality by using the given sequence always.
  - None of the numbers greater than overall optimalities.
- This shows that the agents indeed capture the context from their utilities and behave accordingly.

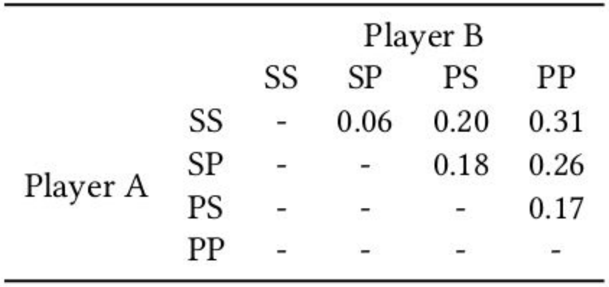
# Interplay between agents



**Varying degrees of selfish/prosocial behavior in agents:**

PP PS SP SS

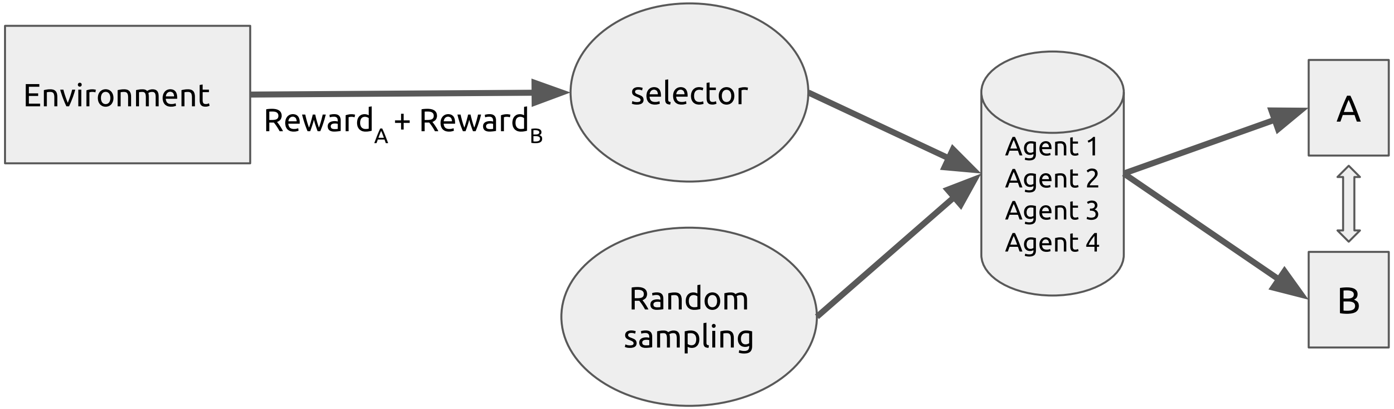
Increasing selfishness



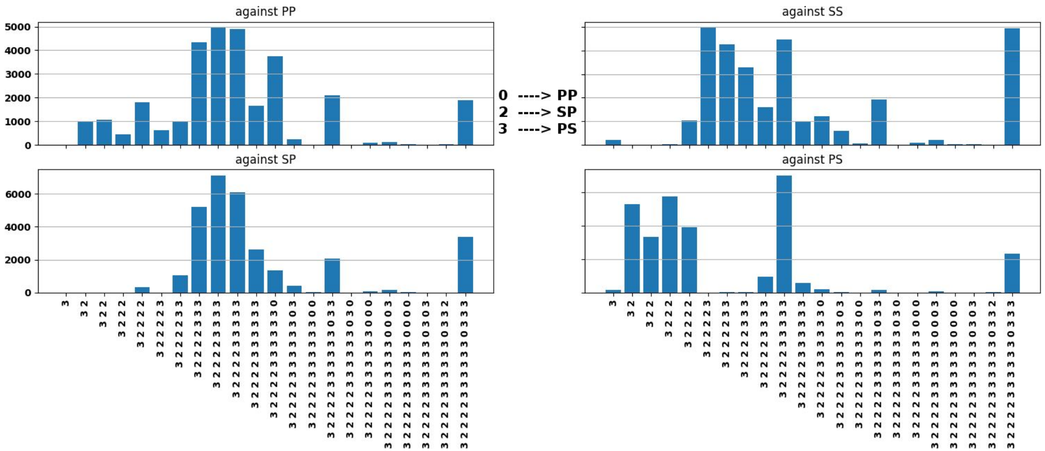
|  |  | Player B |  |  |  |
|---|---|---|---|---|---|
|  |  | SS | SP | PS | PP |
| Player A | SS | - | 0.06 | 0.20 | 0.31 |
|  | SP | - | - | 0.18 | 0.26 |
|  | PS | - | - | - | 0.17 |
|  | PP | - | - | - | - |

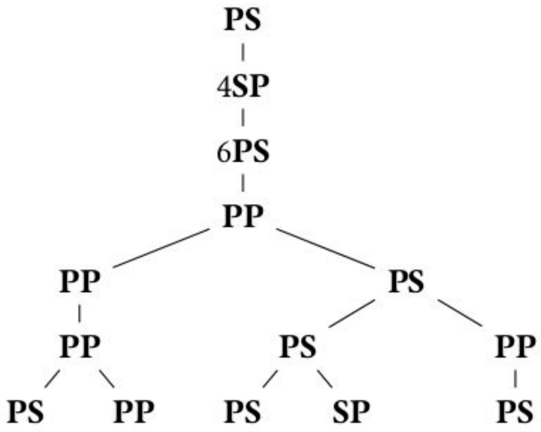
# Mixture of agents (Dynamic Behavior)

- On the surface, the selfish agent seems to be best as it always wins. But it is way too stubborn and won't work well if the other agent is selfish as well.
- We also care about the overall optimality of deals.
- It has been noted previously (Axelrod 1982) that there is no universally best policy to negotiate and that it depends on the nature of the opponent.
    - We try to model one policy that works decently against all agents by using a mixture of agents.
    - We train another agent (selector agent) which selects the behaviour to use at any given step of the negotiation. This agent is also trained using REINFORCE.
    - The reward for this agent is the joint reward of the players.



# Behavior of the Meta Agent





*Selector learns a decision tree:*

- The agent learns just one policy (the simplest) which works against all agents.
- We know that it is difficult for an agent to decipher the behavior of the opponent until after a few moves, hence it makes sense to learn just one policy which works well at any stage.

# Human Evaluation

- To test the performance of the agents against humans:
  - A total of 38 human players negotiated for 3 rounds of negotiation against all 5 agents.
  - Each agent played a total of 114 negotiation games against humans

| Agent | Dialog Length | Agreement Rate (%) | Optimality Rate (%) | Agent score | Human score | Agent won (%) | Human won (%) | Tied (%) |
|---|---|---|---|---|---|---|---|---|
| **PP** | 15.07 | 87.38 | 70.87 | 0.58 | 0.62 | 36.67 | 51.11 | 12.22 |
| **SS** | 19.56 | 73.79 | 60.20 | 0.58 | 0.44 | 60.53 | 21.05 | 18.42 |
| **PS** | 13.57 | 92.93 | 66.67 | 0.57 | 0.57 | 40.22 | 52.17 | 7.61 |
| **SP** | 21.75 | 72.28 | 59.41 | 0.61 | 0.39 | 68.49 | 20.55 | 10.96 |
| **Meta** | 16.78 | 88.30 | 56.40 | 0.57 | 0.56 | 46.99 | 44.58 | 8.43 |

- With the meta agent, humans win an almost equal number of times as the meta agent.
- This proves that we have been somewhat successful in emulating human behavior through our meta agent.

# Future work

- Use Reinforcement Learning to learn hyperparameters involved in the proposal curves.[2]
- Train agents to ground their communication in natural language (NL) while negotiating by making them perform a parallel NL task.[3]
- Analysis and complexity of more than two parties using Reinforcement Learning.

[2] Mukun Cao, Xudong Luo, Xin Robert Luo, and Xiaopei Dai. 2015. Automated negotiation for e-commerce decision making: A goal deliberated agent architecture for multi-strategy selection. Decision Support Systems 73 (2015), 1–14.

[3] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2016. Multi agent cooperation and the emergence of (natural) language.

# Thank you!!