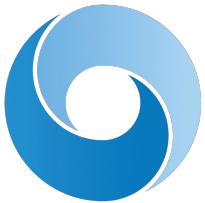
Iterative Empirical Game Solving via Single Policy Best Response

Max Olan Smith[1], Thomas Anthony[2], Michael P. Wellman[1]
[1] Michigan, [2] Google DeepMind

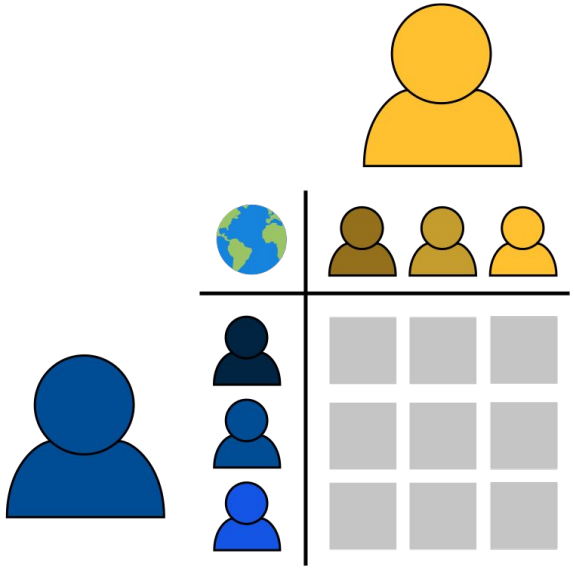


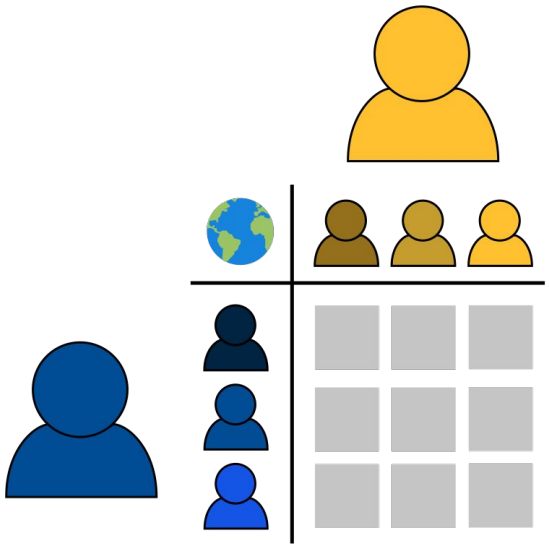
Problem:
Reduce amount of training required in multiagent systems.

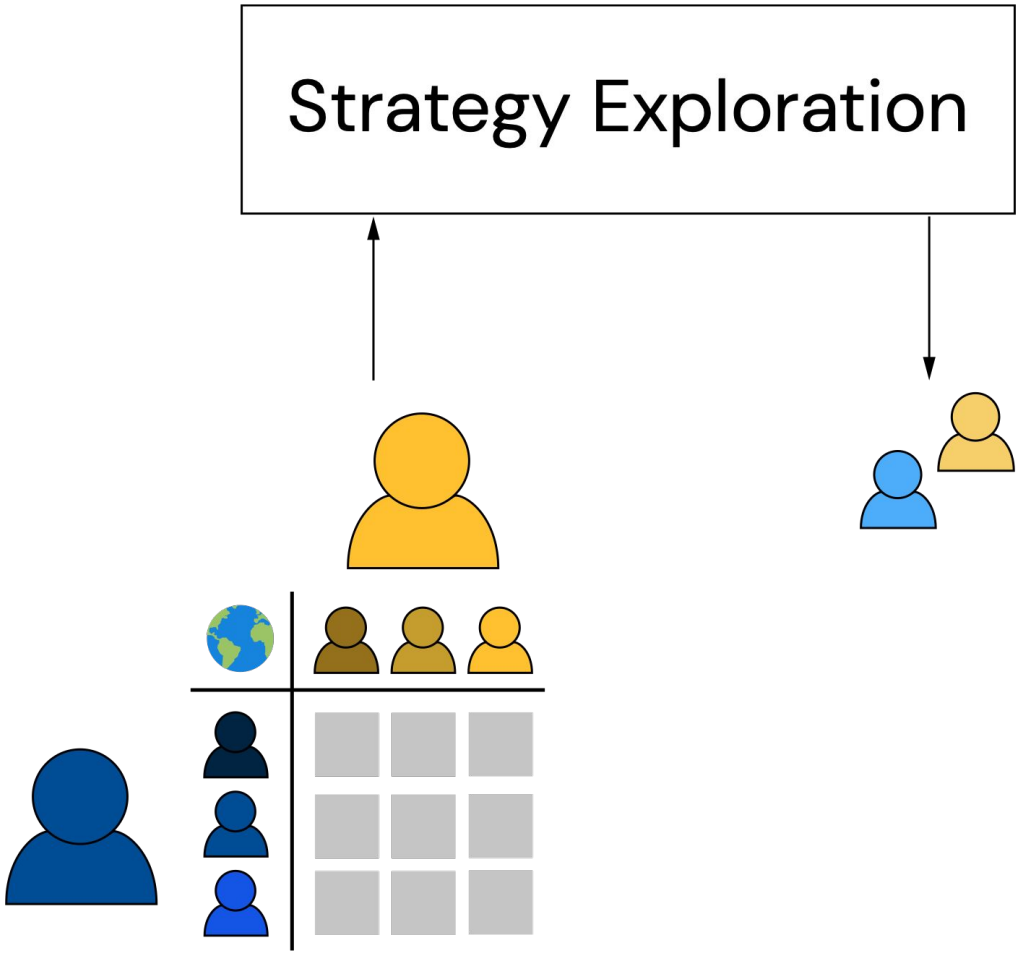
Real Game

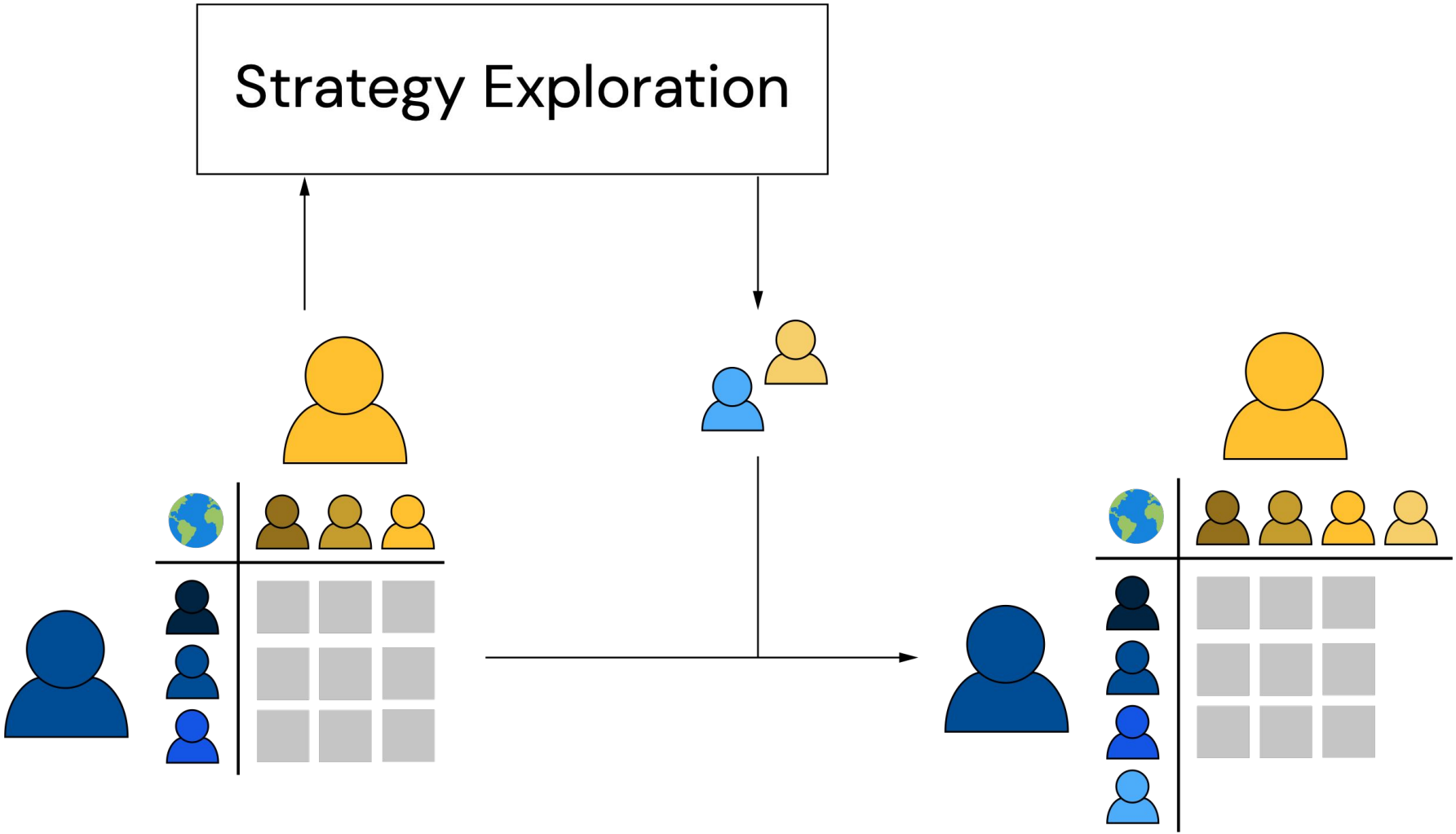
Problem:
Reduce amount of training required in multiagent systems.

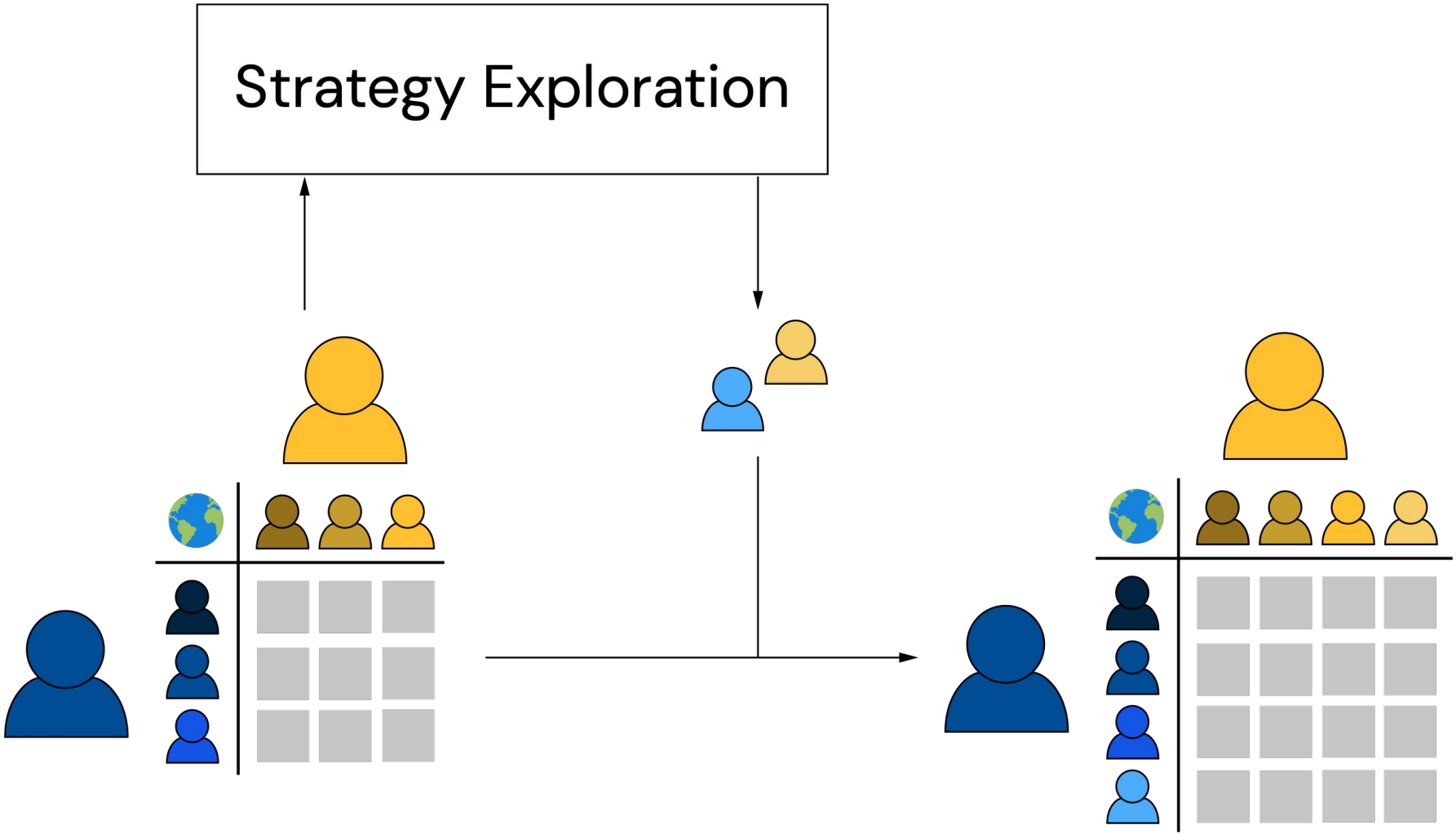
Empirical Game

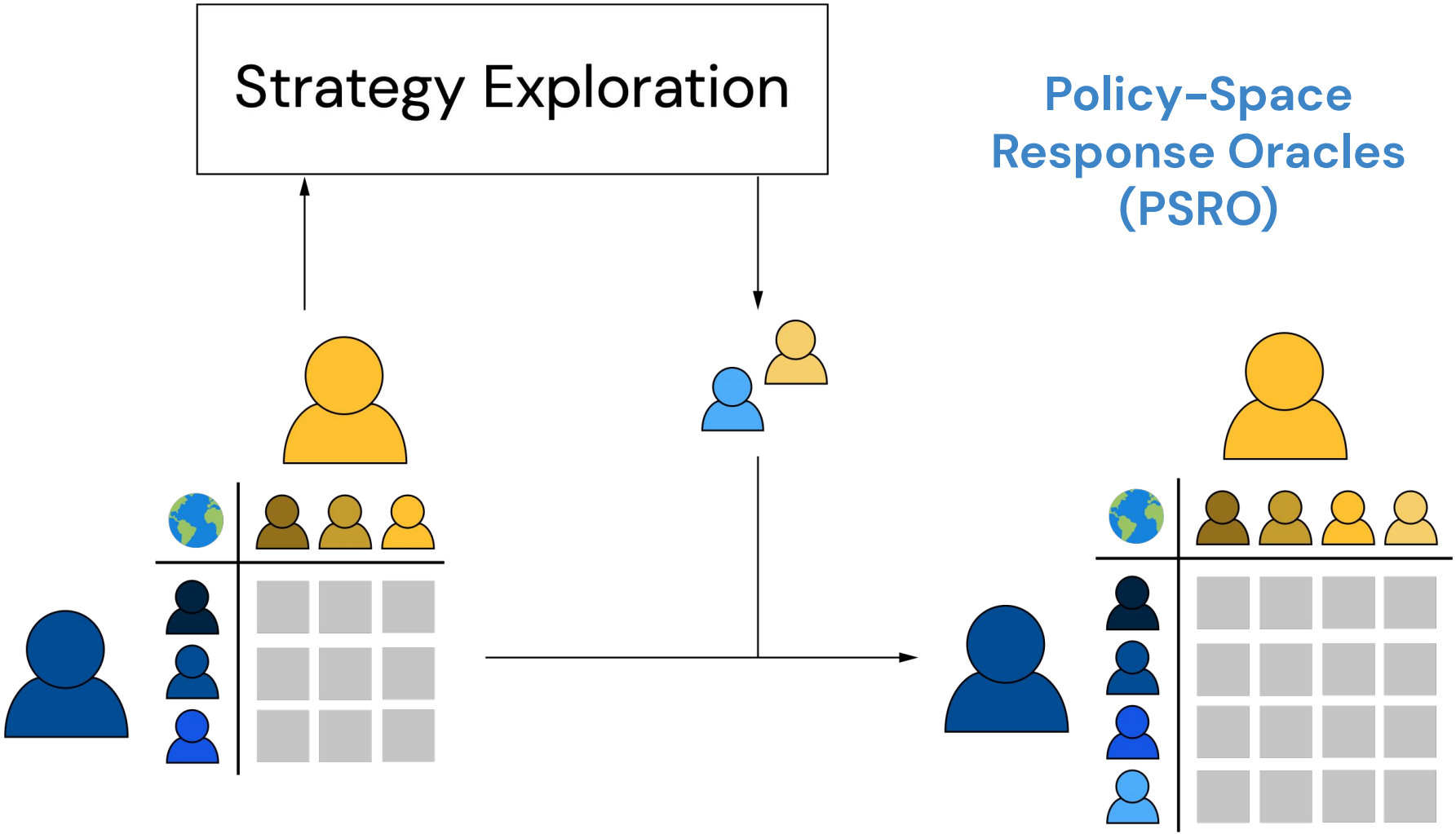





Strategy Exploration

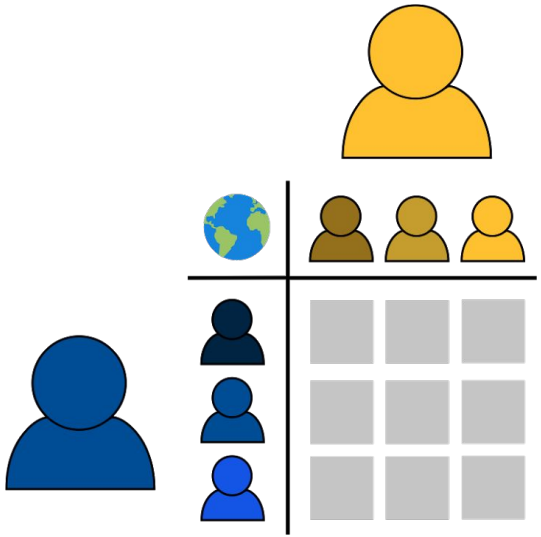

Strategy Exploration


Strategy Exploration


Strategy Exploration
Policy-Space
Response Oracles
(PSRO)

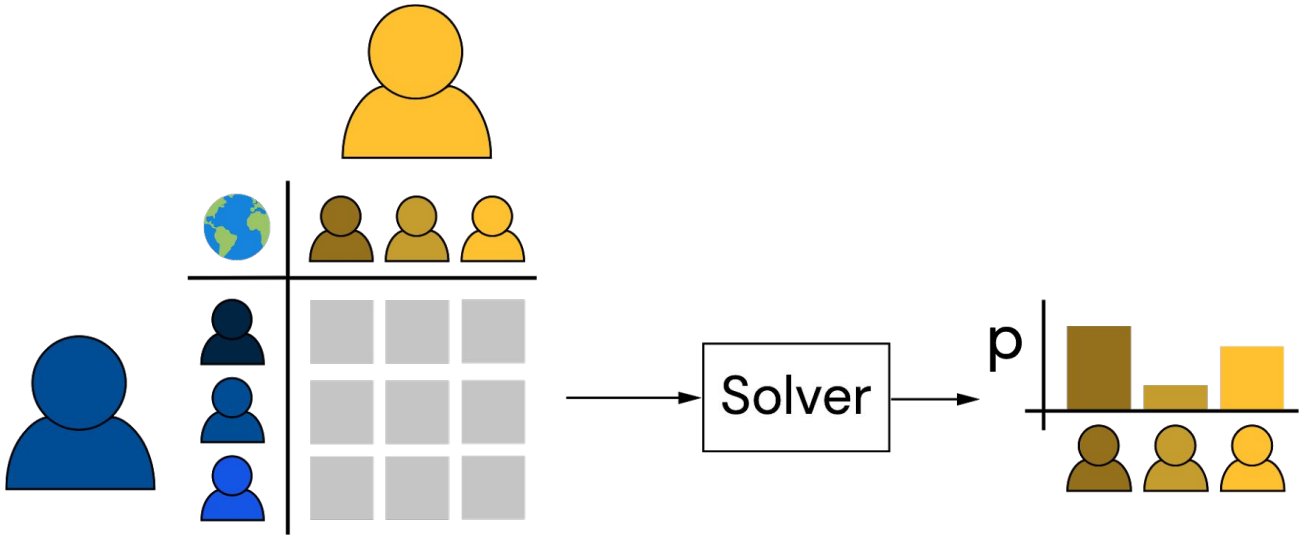
Strategy Exploration

Which policies to add to the empirical game?



Strategy Exploration

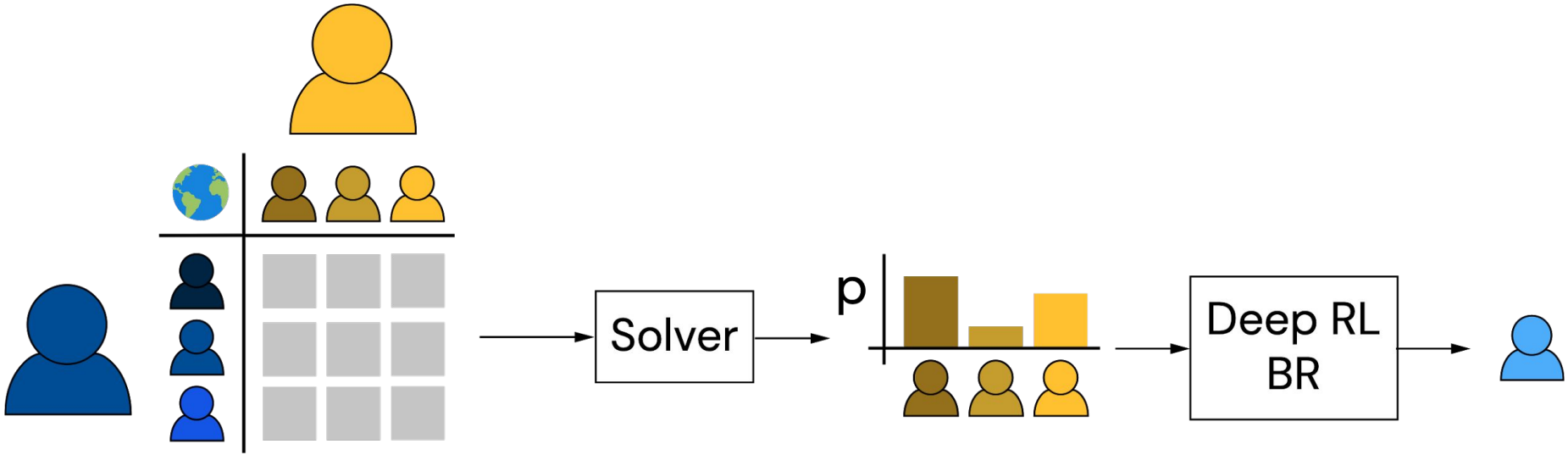
1. Solution:
Solve current empirical-game.



Strategy Exploration

1. Solution:
Solve current empirical-game.

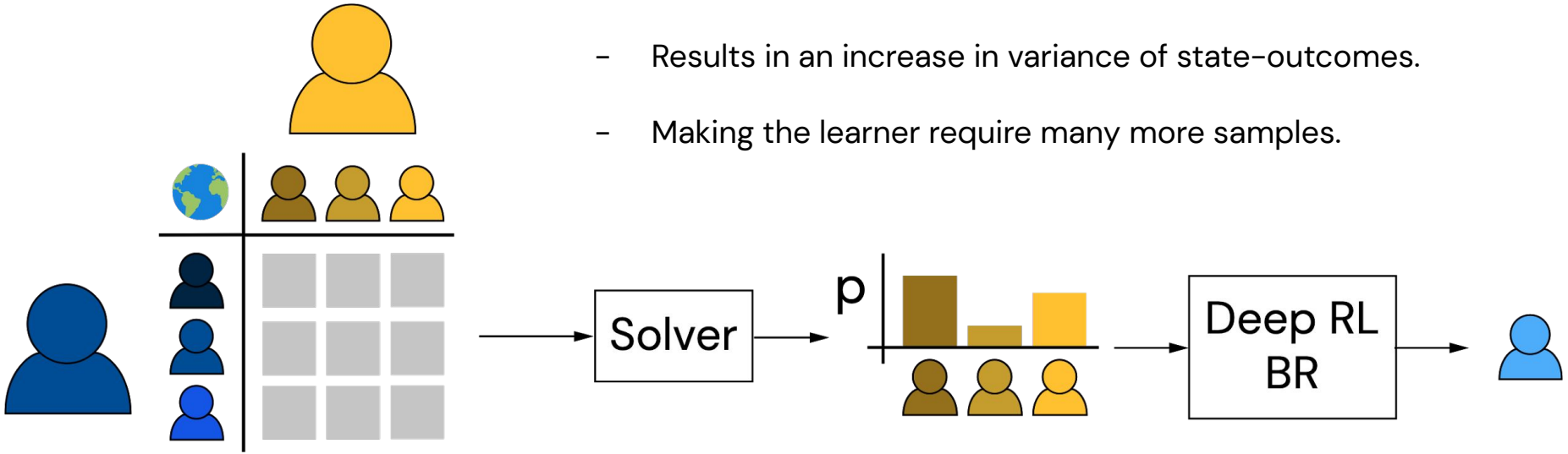
2. Best Response:
Add best-responses for each Player to the current solution.



Strategy Exploration

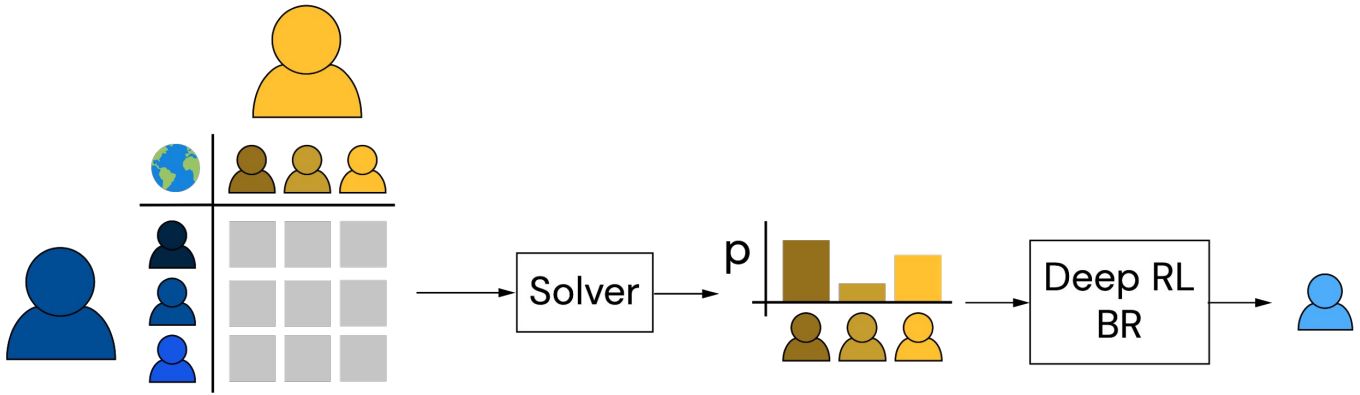
During training:

- Opponent sampled at start of episode.
- Agent does not observe their current opponent.
- Results in an increase in variance of state-outcomes.
- Making the learner require many more samples.



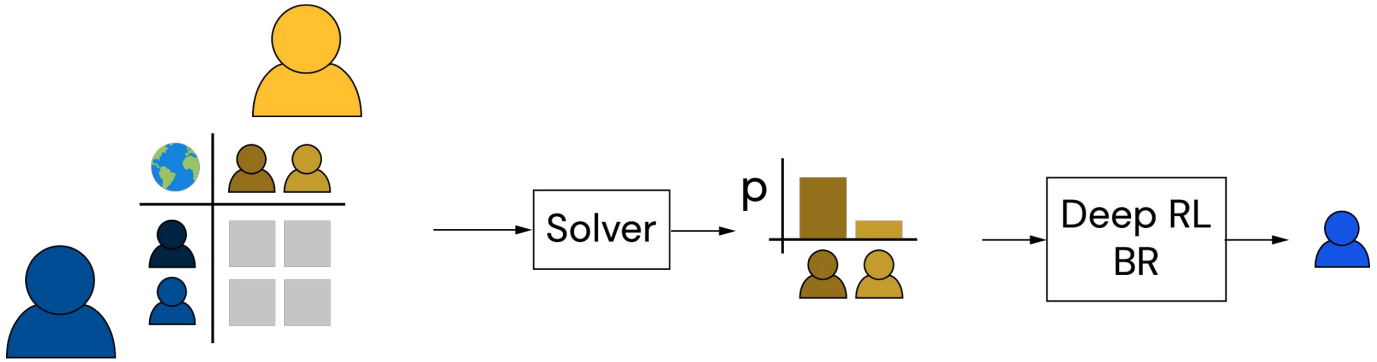
Strategy Exploration

Epoch 3:

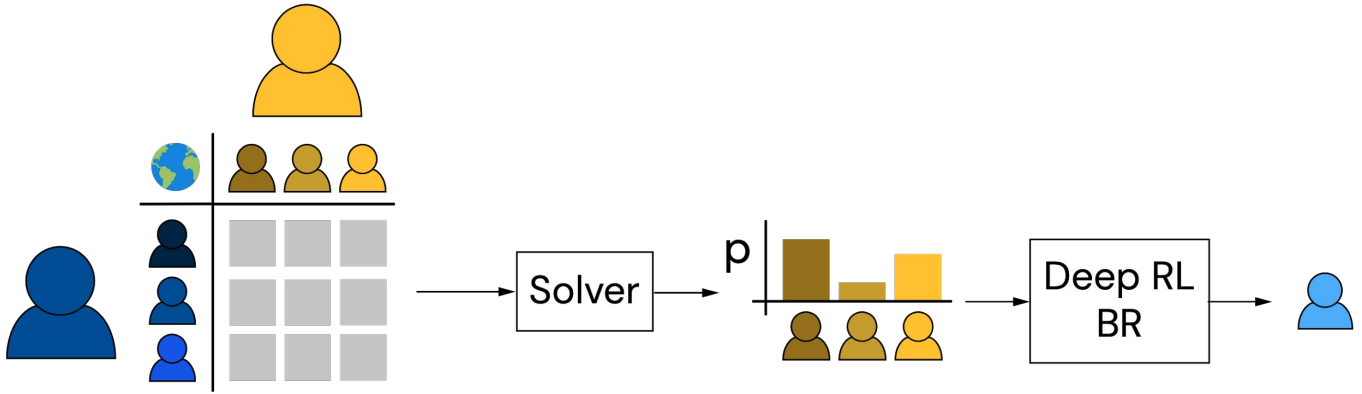


Strategy Exploration

Epoch 2:

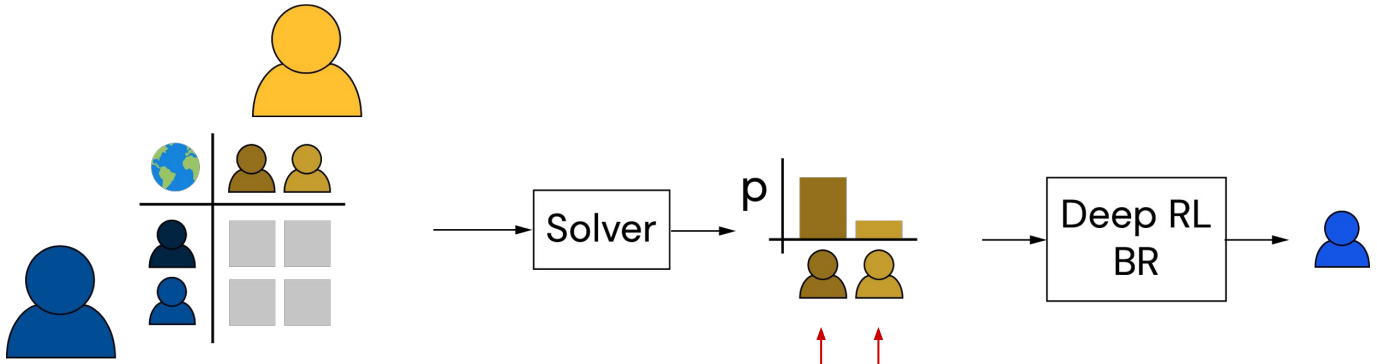


Epoch 3:

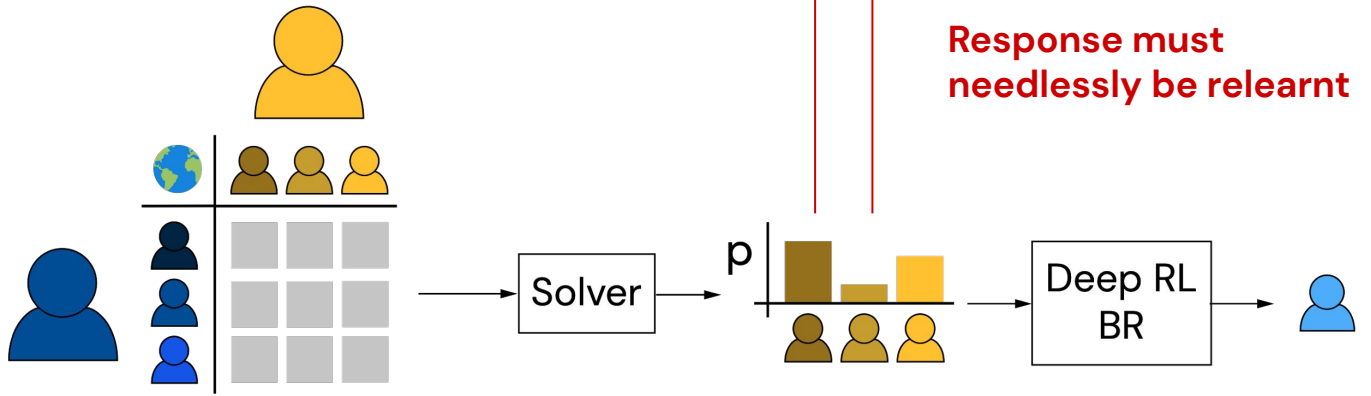


Strategy Exploration

Epoch 2:



Epoch 3:

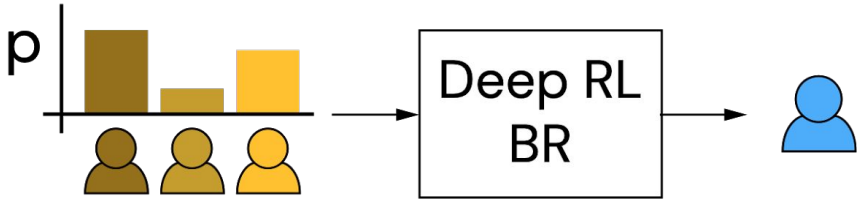


Response must needlessly be relearnt

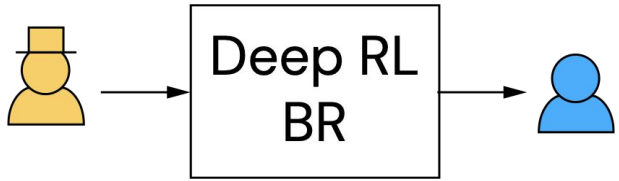
Contributions

- Replace mixed-strategy opponent with a single opponent policy.
- Focus training on single most salient opponent policy.
- Reducing variance during learning.

Current Approach



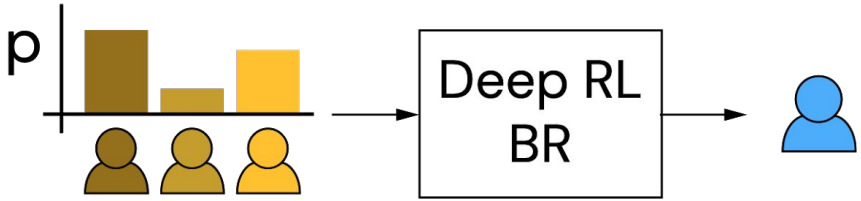
Proposed Approach



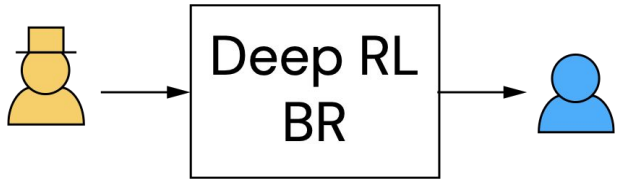
Contributions

- Replace mixed-strategy opponent with a single opponent.
- Focus training on single most salient opponent.
- Reducing variance during learning.
- Two new algorithms differing choice of new opponent:
  - **Mixed-Oracles**
  - **Mixed-Opponents**

Current Approach



Proposed Approach

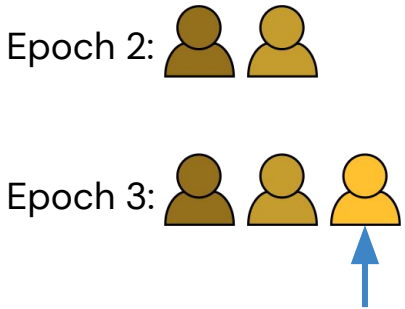


Mixed-Oracles

Insight: each player adds one new policy each epoch.

Idea:

- Train best-response to new policy.
- Transfer best-responses from older policies.

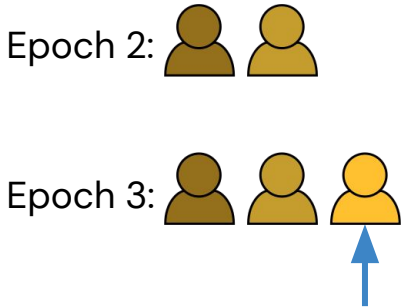


Mixed-Oracles

Insight: each player adds one new policy each epoch.

Idea:

- Train best-response to new policy.
- Transfer best-responses from older policies.

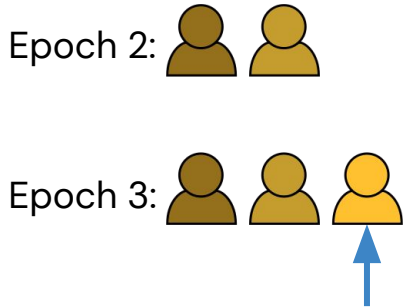


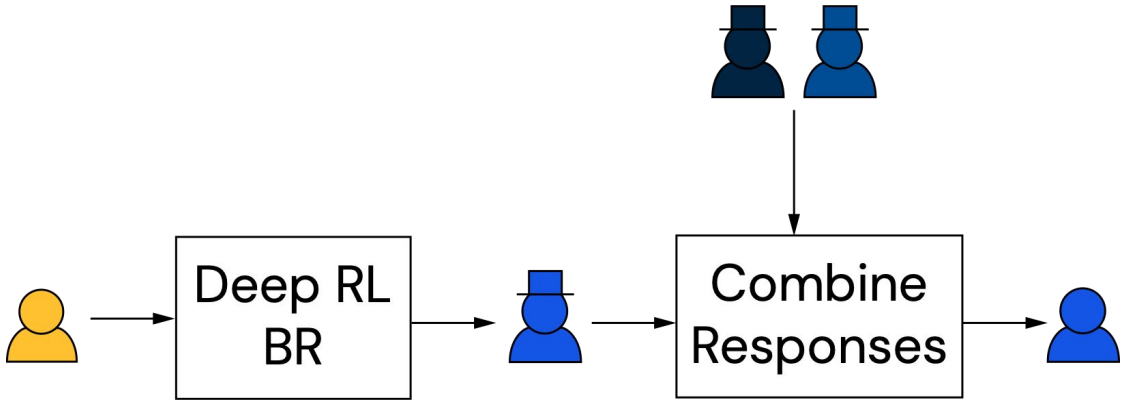
Mixed-Oracles

Insight: each player adds one new policy each epoch.

Idea:

- Train best-response to new policy.
- Transfer best-responses from older policies.

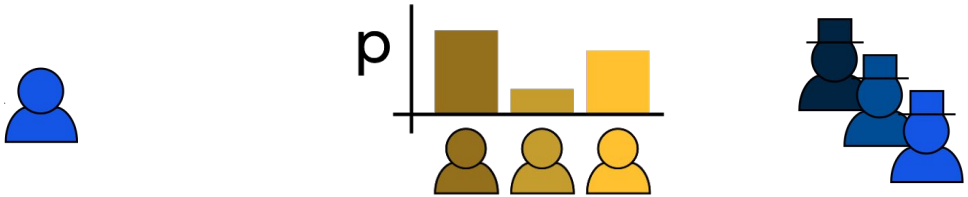




Combine Responses

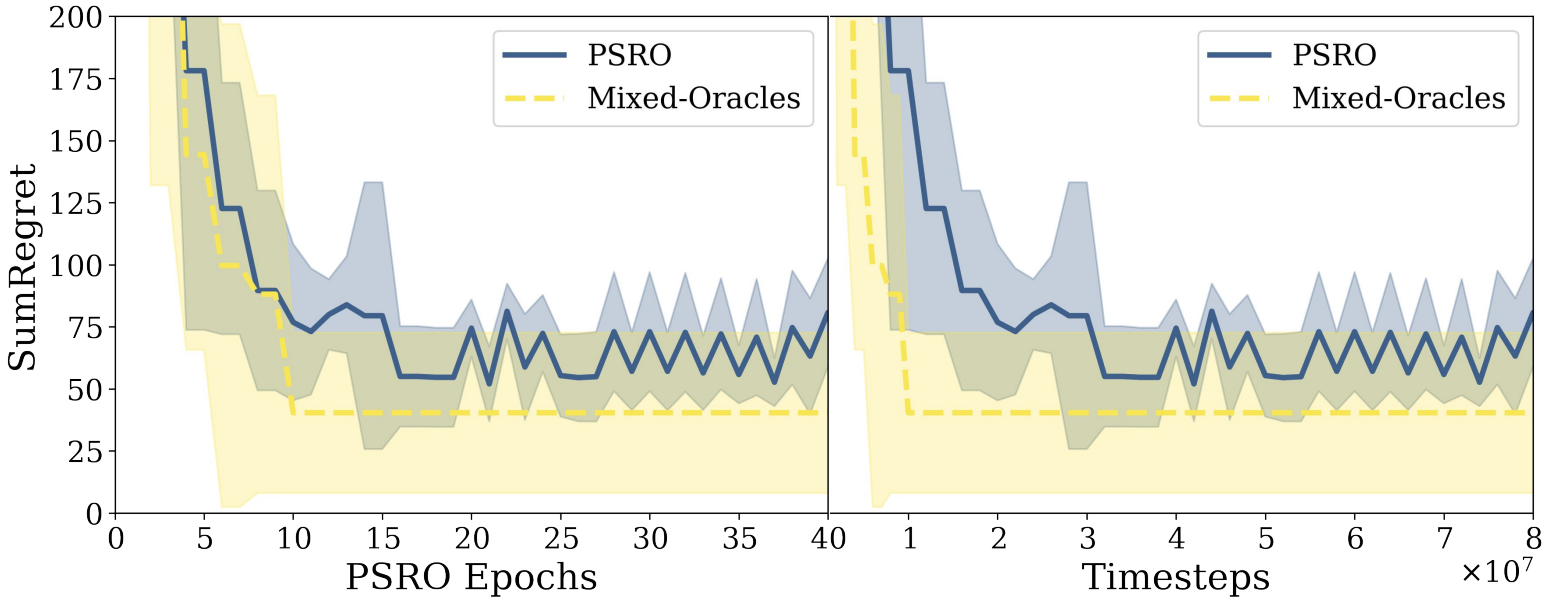
- Using our previous work: **Q-Mixing**.
  - Get approximate best-response to any mixed-strategy,
  - By averaging Q-values of pure-strategy best-responses.

BR Mixed-Strategy $= \sum_{\pi_{-i}}$ Opponent Likelihood BR Pure-Strategy



$$Q_{\pi_i}(o_i, a_i \mid \sigma_{-i}) = \sum_{\pi_{-i}} \psi_i(\pi_{-i} \mid o_i, \sigma_{-i}) Q_{\pi_i}(o_i, a_i \mid \pi_{-i})$$

Mixed-Oracles: Gathering-Small



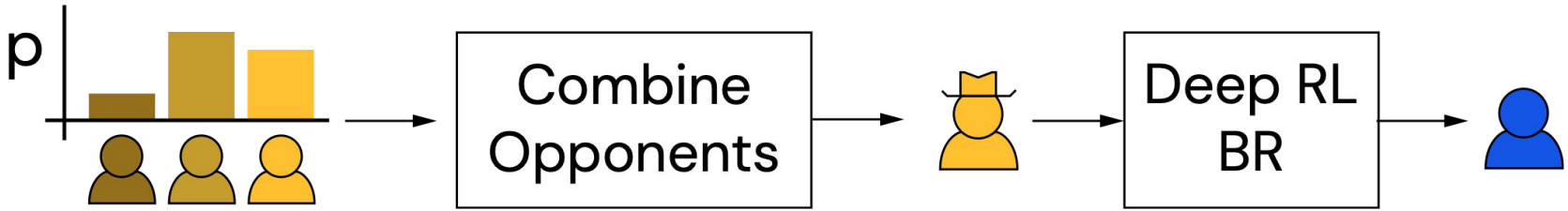
Mixed-Opponents

- The opponent's mixed-strategy may not be the most salient opponent.
- Explore strategy-space more efficiently, reduce number of epochs.

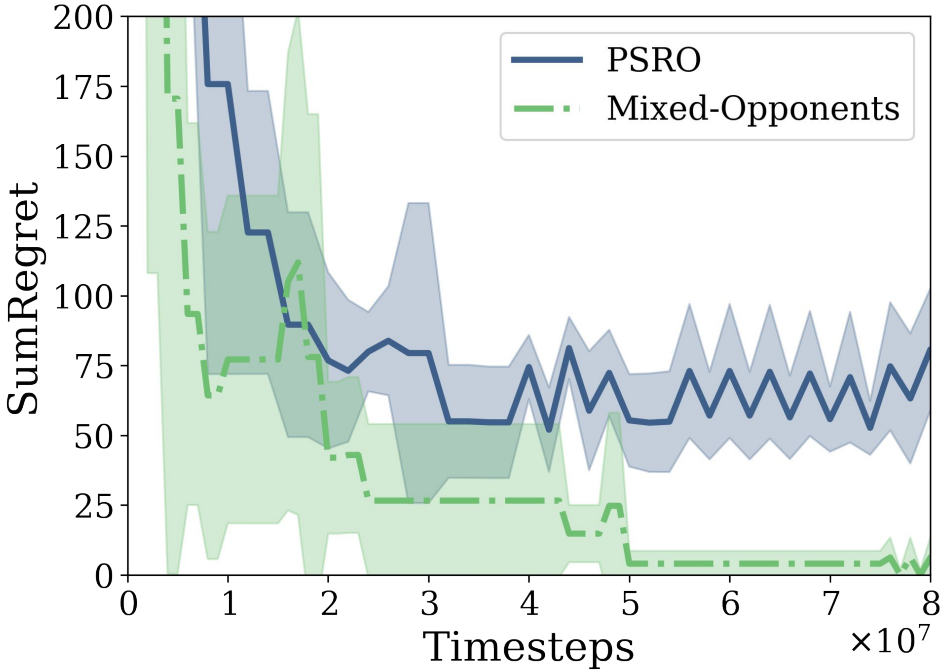
Mixed-Opponents

- The opponent's new policy may not be the most salient opponent.
- Explore strategy-space more efficiently, reduce number of epochs.

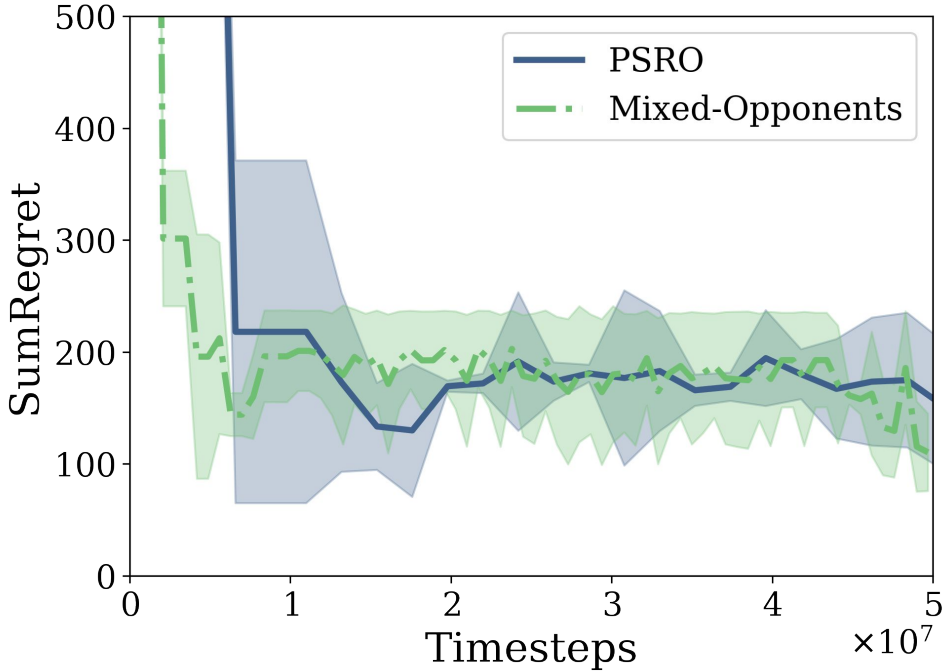
- **Idea:** combine opponent policies by averaging their Q-values.
  - Considers each policy's value for each action respectively.
  - Giving rise to a unique greedy policy.



Gathering–Small 2–Player

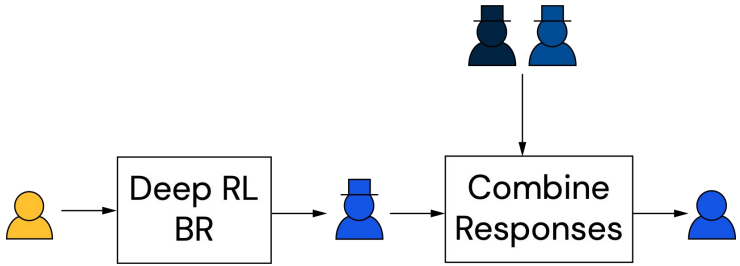


Gathering–Open 3–Player

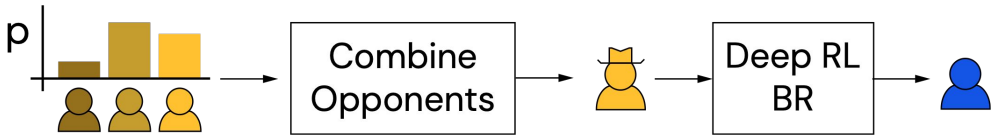


Conclusion

- Replace mixed-strategy objective with pure-strategy objective in PSRO.
- Offer reductions in training time and may converge to better solutions.



Mixed-Oracles: transfer knowledge about past opponents.



Mixed-Opponents: aggregate opponents into single new policy.

Hyperparameter Ablation (Gathering-Small):