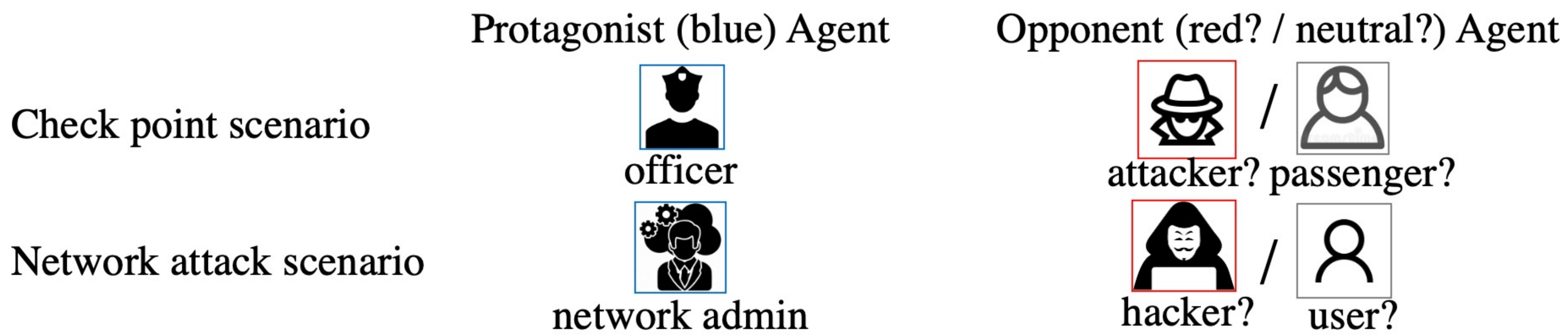
Robust Opponent Modeling via Adversarial Ensemble Reinforcement Learning with Uncertain Opponent Types

Macheng Shen and Jonathan P. How
Aerospace Controls Laboratory, Massachusetts Institute of Technology

Motivating Examples



- Protagonist agent must infer opponent type to make optimal decision
- Making wrong decision leads to catastrophic consequences

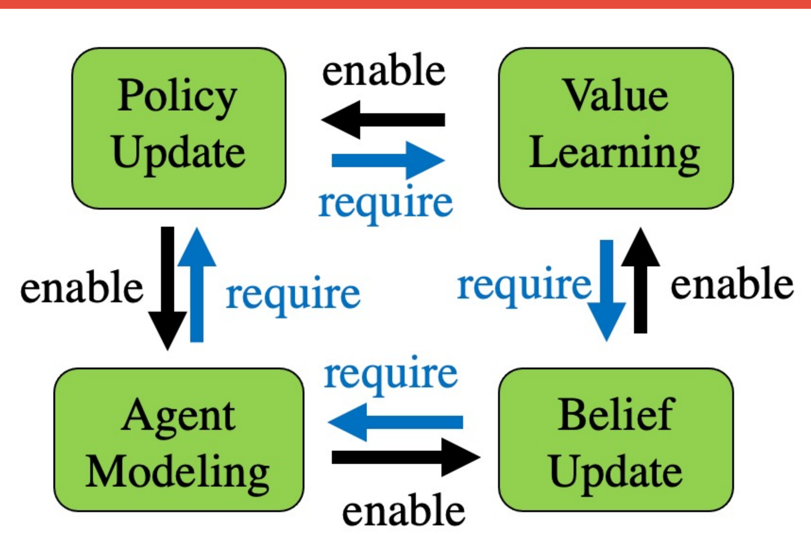
Problem Statement & Formulation

- We are interested in multiagent scenarios decision-making problem **with uncertain opponent types** ➔ critical information for making right decision
- Decision-making framework: Bayesian Game

$$\langle \mathcal{I}, \langle \mathcal{S}, \mathcal{H} \rangle, \{b^0\}, \{\mathcal{A}_i\}, \{\mathcal{O}_i\}, \mathcal{P}, \{R_i\} \rangle$$

  - $\mathcal{I}$: Information state space
  - $\langle \mathcal{S}, \mathcal{H} \rangle$: Joint space of state and **agent type**
  - $\{b^0\}$: Initial belief over agent type
  - $\{\mathcal{A}_i\}$: Joint action space
  - $\{\mathcal{O}_i\}$: Joint observation space
  - $\mathcal{P}$: State transition probability
  - $\{R_i\}$: Reward function, depends on **both state and agent type**
- Objective: $V^{\pi}(b_0) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{s^t \sim p_0(s^t, h^t), a^t \sim \pi(b^t)} [r(s^t, a^t, h^t)]$
- Subject to: $b^{t+1} = \textbf{Belief Update}(b^t, o^t)$

Challenges



- **Coupled** belief & policy update
- **Information asymmetry**: red knows blue, blue does not know red → Incentive for red to deceive → Difficult to model

Related works

- Plan recognition [1; 2]
  - Type inference using a pre-defined set of agent models
  - Limitation: **Inaccurate modeling** → biased belief
- (Planning-based) Multiagent reasoning
  - Game-theoretic agent modeling, Bayesian-Nash Equilibrium
  - Limitation: **Poor scalability**, feasibility restricted to matrix games [3], two step games [4]

Our approaches

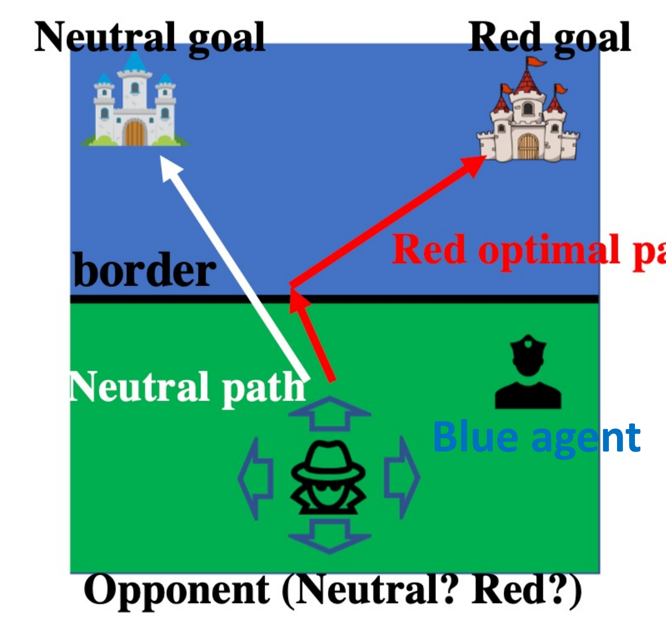
- **Game-theoretic opponent modeling** based on MARL
  - Simultaneously model both agents
  - **Capture strategic interaction** between agents
  - ➢ Improve modeling accuracy, better scalability
- Diversity-driven **ensemble opponent modeling**
  - Ensemble training $J(\pi_i) = \mathbb{E}_{\substack{k \sim \text{unif}(1,K), \\ a_i \sim \pi_i(b_i), \\ a_{-i} \sim \pi_{-i}^{(k)}}} \left[\sum_{t=0}^{\infty} \gamma^t r_i(b_i, \boldsymbol{a}) \right]$
  - Diversity-driven evolutionary optimization
  - ➢ Improved robustness against adversary
- Exact **belief update & belief-space reward**

$$b_i^t \propto \mathbb{E}_{\boldsymbol{a}^t \sim \boldsymbol{\pi}(\bar{o}|\boldsymbol{h})} [\mathcal{P}^O(o_i^t | \boldsymbol{a}^t, s^t)] \int p(s^t | s^{t-1}, h^t) b_i^{t-1} ds^{t-1}$$

Agent Model Observation Prob. State Transition

  - ➢ Lower variance → Stable training
- Contribution: Effective framework for planning under opponent type uncertainty

Experiment



- Security game
  - Opponent type hidden to blue
  - Blue infers opponent type
  - Blue cannot pass the border
  - Blue can tag the opponent

Figure 1: Blue can tag the opponent and get reward if tagged red, or get penalty if tagged neutral. Consequence: red pretends to be neutral before passing the border (deception)

Results

Q1: Is ensemble training necessary?

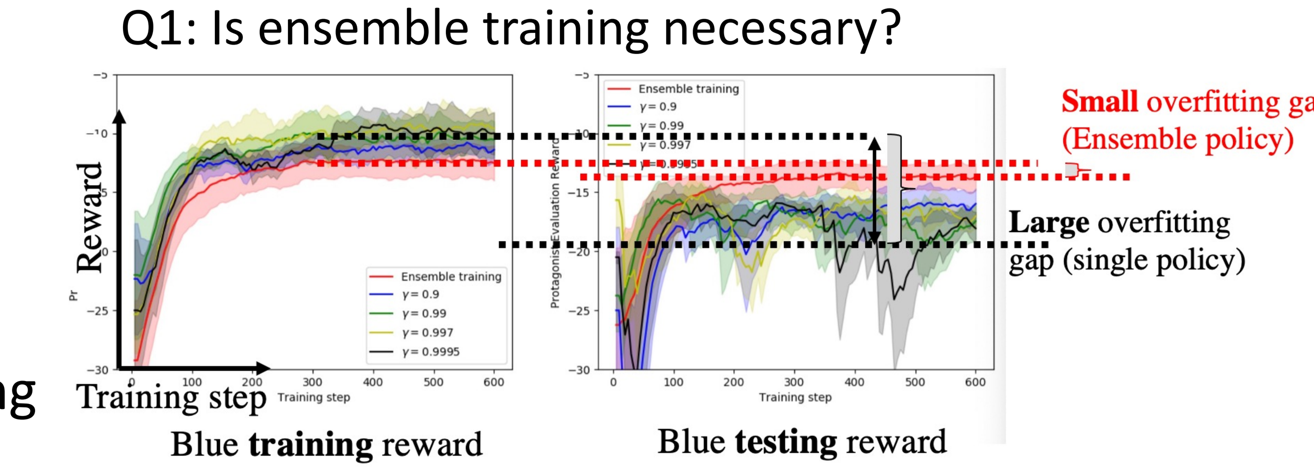


Figure 2: Ensemble training significantly reduces the generalization gap between training and testing.

Results

Q2: Does belief update/belief-space reward help learning?

Learning setting	Protagonist(Blue) reward	Adversary(Red) reward
belief-space policy, with ensemble	**-14.4±1.49**	**-83.0±17.0**
RNN policy, with ensemble	-17.7±1.9 ↓3.3	-66.2±13.8 ↑16.8
belief-space policy, w/o ensemble	-16.5±1.1	-58.6±24.9
RNN policy, w/o EO & CE	-16.8±3.1	-49.4±6.6

Table 1: Comparison between belief space policy and recurrent policy. Belief-space policy achieves higher blue reward and lower red reward, which is consistent across settings. This indicates that belief-space reward indeed helps learning stronger blue policy.

Q3: Is game-theoretic opponent modeling necessary?

Metrics/Approach	MDP Agent-Single	Game theoretic-Single	Game theoretic-Ensemble
Precision t=10	0.57	0.56	**0.82** ↑0.26
Precision t=20	0.60	0.59	**0.89** ↑0.30
Recall t=10	0.12	0.34 ↑0.22	**0.68** ↑0.34
Recall t=20	0.06	0.26 ↑0.20	**0.74** ↑0.48
Protagonist (blue) reward	-19.44	-17.55	**-14.4**
Adversary (red) reward	-50.48	-58.19	**-83.0**

Table 2: Precision and recall of opponent type inference. The recall of MDP agent model is quite low. Game-theoretic modeling with single policy improves the recall, but it is still not high enough, need ensemble to avoid overfitting.

Conclusions

- We proposed an effective framework for planning under opponent type uncertainty that
  - Outperforms single-agent modeling
  - Achieves high type inference accuracy
  - Robust to previously unseen adversaries

References

[1] Fagan, Michael, and Pádraig Cunningham. "Case-based plan recognition in computer games." International Conference on Case-Based Reasoning. Springer, Berlin, Heidelberg, 2003.
[2] Sohrabi, Shirin, Anton V. Riabov, and Octavian Udrea. "Plan Recognition as Planning Revisited." IJCAI. 2016.
[3] Huang, Linan, and Quanyan Zhu. "Dynamic bayesian games for adversarial and defensive cyber deception." Autonomous cyber deception. Springer, Cham, 2019. 75-97.
[4] Nguyen, Thanh H., et al. "Deception in finitely repeated security games." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.