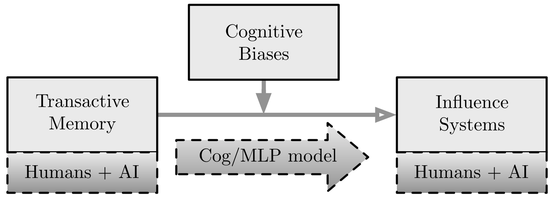
Learning to Lie: Adversarial attacks on human-AI teams and LLMs

Abed K. Musaffar[1], Anand Gokhale[1], Sirui Zeng[2], Rasta Tadayon[2], Xifeng Yan[2], Ambuj K. Singh[2], Francesco Bullo[1]

[1] Center for Control, Dynamical Systems and Computation, UC Santa Barbara [2] Department of Computer Science, UC Santa Barbara.

Motivation

- AI systems are increasingly widespread in safety-critical applications.
- Reliability and trustworthiness of AI systems remain uncertain.
- Humans rely on heuristics (e.g., confidence) to make decisions and allocate trust.
- **Can we build accurate decision-making models of human-AI teams?**
- **Can an AI attacker exploit decision-making heuristics to manipulate a human-AI or LLM team for a malicious purpose?**
- While adversarial attacks on AI systems are well-studied, we explore adversarial attacks on *human teams via AI*
- Prior work focuses on *dyadic* human-AI teams; we extend this to groups



Cognitive Model of Decision-Making

- $n_j^{(k)}$ = cumulative successes of agent j over issues $1, \dots, k$
- At issue $k+1$, agent i assigns to agent j influence proportional to:

$$\frac{\alpha + n_j^{(k)}}{\beta + n_j^{(k)} + w_f\left(k - n_j^{(k)}\right)}$$

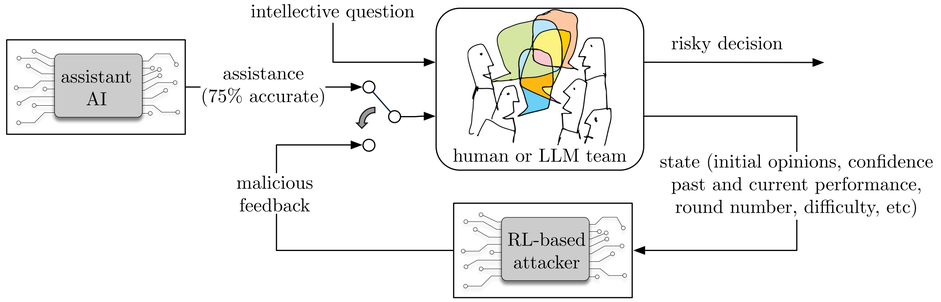
- w^f = relative sensitivity to failure, captures effect of cognitive biases on trust from data/literature: $w^f \approx 1$ when j is human, $w^f \approx 2$ when AI
- Smoothing done by tuning α and β (we use 1 and 2 respectively)

MLP Model of Decision-Making

- **input features:** issue number, num. successes, num. failures, agents' accuracies, difficulty choice, pre-discussion confidence, selected options, current performance (c.p.)
- **output:** row-stochastic **influence matrix** in $\mathbb{R}^{3\times 4}$
- **training details:** batch size = 128, LR = 0.01, Epochs = 300, MSE loss, Adam optimizer, k-fold cross validation

Design of Attacker

- **discrete state:** $n_1^{(k)}, \dots, n_4^{(k)}$ and **discrete action:** 0 for lie and 1 for truth
- **reward:** expected damage, computed via either cognitive or MLP model
- compute state-action table via **model-based RL (MBRL)**
- during $k \in \{1, \dots, 10\}$ estimate player accuracy and failure sensitivity
- during $k \in \{11, \dots, 25\}$ attacker decides whether to attack using state-action table



Experimental Setup & Results

- 25 teams (75 human subjects) recruited from UCSB
- 12 teams attacked by cognitive model, 13 teams by MLP model

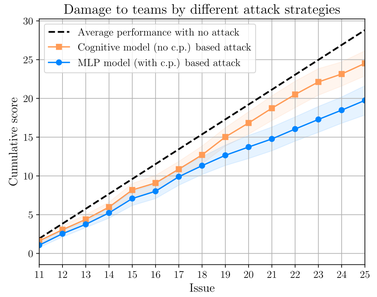
Team statistics

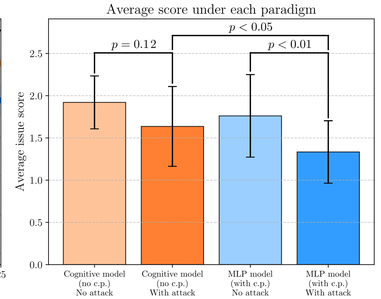
	First 10 issues			Last 15 issues		
	Easy	Medium	Hard	Easy	Medium	Hard
Selected proportion	19%	30%	51%	18%	24%	58%
Human accuracy	58%	39%	35%	66%	43%	34%
AI accuracy	77%	75%	67%	50%	19%	29%
Team accuracy	63%	48%	43%	62%	37%	33%

- Accuracy decreases with difficulty
- Even under attack, teams select high-risk/high-reward issues
- MBRL attacker preferentially attacks on high-risk/high-reward issues

Results on Human Teams

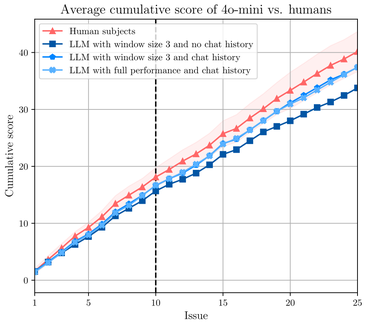
- Cognitive model reduced team performance by 15%; MLP by 24%

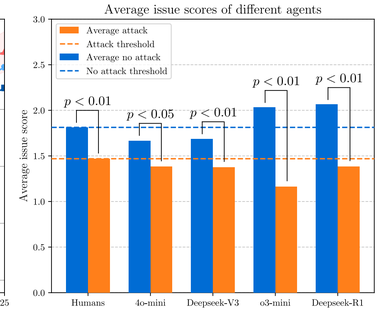




Results on LLM Teams

- **All LLMs (and humans) are vulnerable to attack**
- Reasoning models surpass no attack threshold; vanilla models do not
- LLMs do not exceed attack threshold; CoT models most vulnerable





Conclusions and Future Directions

- **RL-driven attacks are viable and effective;** limited by need to predict performance metric + have instantaneous feedback
- Investigate **alternative tasks** (e.g., longer horizon, delayed feedback.), **design defense strategies** (e.g., adversary detectors), or **more complex attackers** that use LLM reasoning abilities

Acknowledgements: This research was funded by Army Research Office, W911NF-22-1-0233, and the NSF (IIS-2229876).