Deceptive Sequential Decision-Making via Regularized Policy Optimization

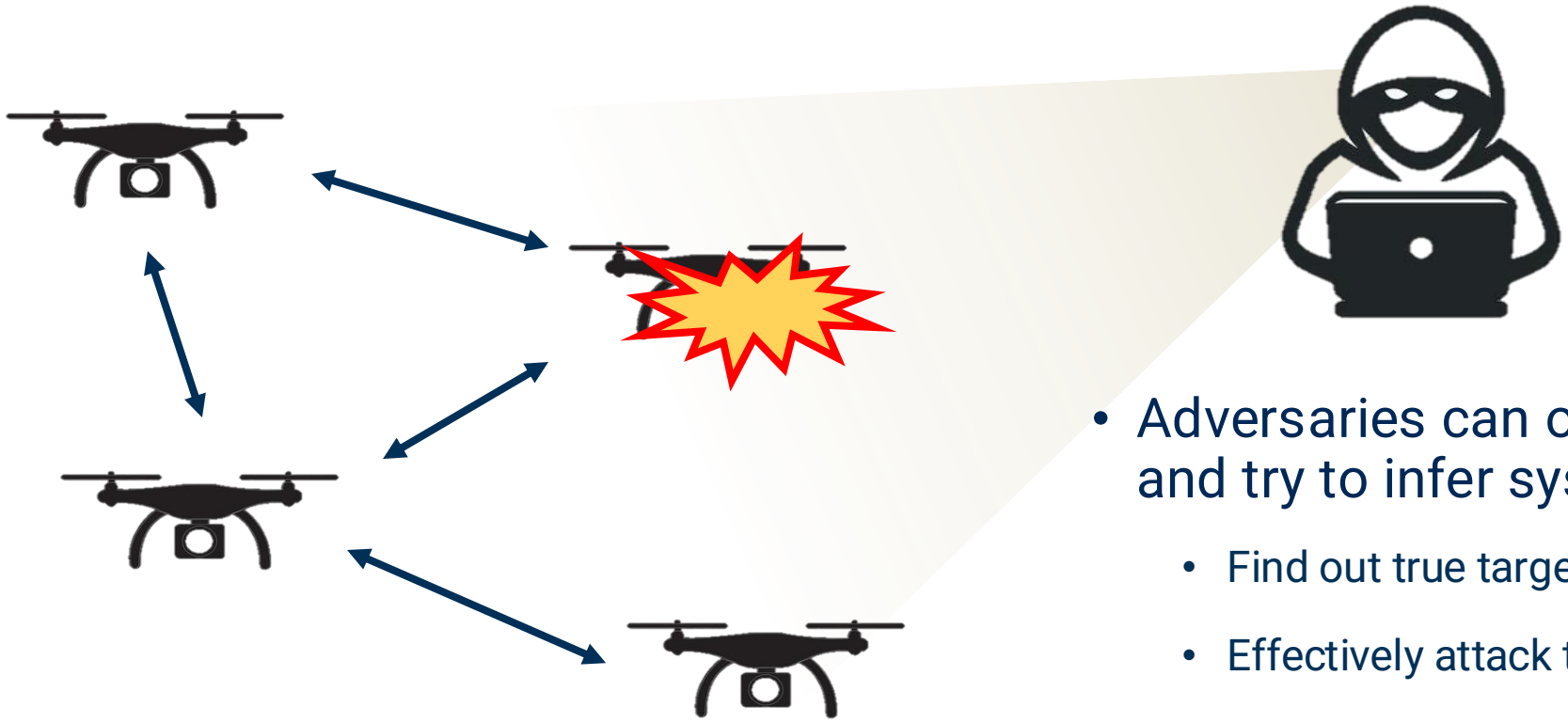
Yerin Kim, Alexander Benvenuti, Bo Chen, Mustafa Karabag,
Abhishek Kulkarni, Nathaniel D. Bastian, Ufuk Topcu, Matthew Hale

Affiliations: Georgia Institute of Technology, University of Texas at Austin, United States Military Academy West Point

Georgia Tech

Concealing the true objective is crucial

- Systems operating in physical environment are inherently observable.
- A network of drones performing a collaborative task.

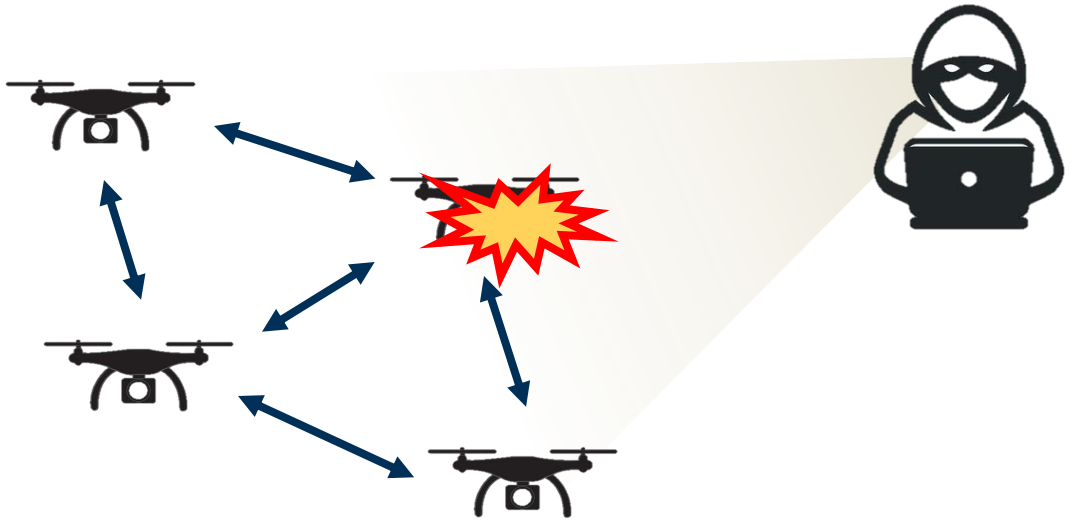


- Adversaries can observe the system and try to infer system objective.
  - Find out true targets
  - Effectively attack the system

Georgia Tech

We want to conceal the objective

- Adversaries attempt to infer the system objective by observation.
- We expect the system to be observed, but we cannot prevent observations.



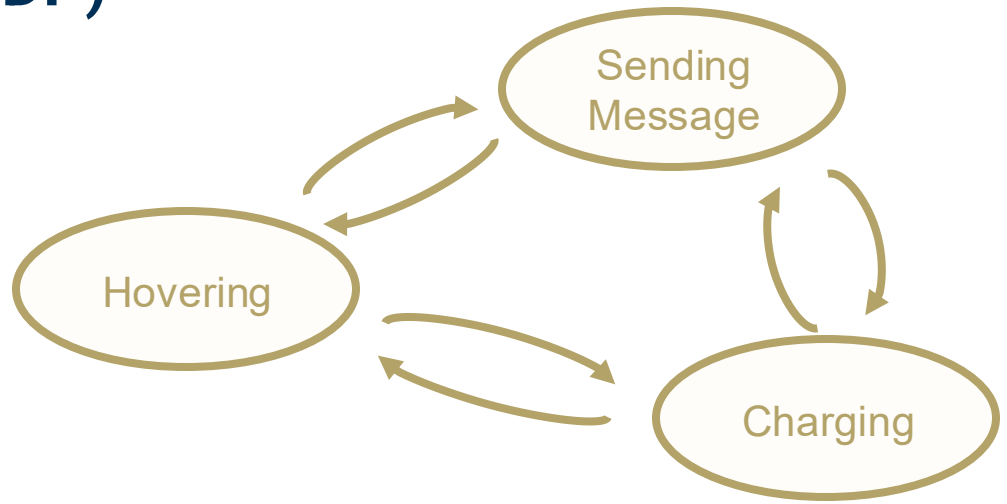
Question 1: How do we conceal the objectives of the system under complete observation?

Question 2: Can we still achieve good performance while concealing the objective?

We used MDP to formulate problem

Goal: Deceive adversary by pretending to have another objective.

- System model: **Markov Decision Process (MDP)**
  - $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{R}, \gamma, \mathcal{T})$
    - $\mathcal{S}$: Set of states
    - $\mathcal{A}$: Set of actions
    - $\mathcal{R}$: Rewards (Reflects the system objective)
    - γ: Discount factor
    - $\mathcal{T}$: Transition matrix



- Adversary model: **Inverse Reinforcement Learning (IRL)**
  - Aims to recover the reward $\mathcal{R}$ based on observed state-action trajectories
  - Rewards are calculated based on the frequency of the state-action visitation

Goal: Deceive adversary by "manipulating state-action visitation frequency".

B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey et al., "Maximum entropy inverse reinforcement learning." in AAAI, vol. 8. Chicago, IL, USA, 2008, pp. 1433–1438

We used dual formulation to solve MDP

- IRL uses state-action visitation frequency to infer the rewards.
- A **policy π** is a choice of action that maximizes accumulated reward.
  - One way to find a policy is linear programming.
  - Dual formulation of problem (No deception yet)

Occupancy measure $x(s,a)$
↔ State-action visitation frequency

$$\max_{x} \quad \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} r(s,a)x(s,a)$$

← Maximize accumulated reward (reach goals)

$$s.t. \quad \sum_{a\in\mathcal{A}} x(j,a) - \gamma \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \mathcal{T}(s,a,j)x(s,a) = \alpha(j), \qquad \forall j \in \mathcal{S}$$

← Obey the transition probability

$$\sum_{q\in\mathcal{S}_{goal}} \left[\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \mathcal{T}(s,a,q)x(s,a) + 1(q)\right] \geq v_{reach}$$

← Ensure frequent visitation to goals

$$x(s,a) \geq 0, \qquad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}.$$

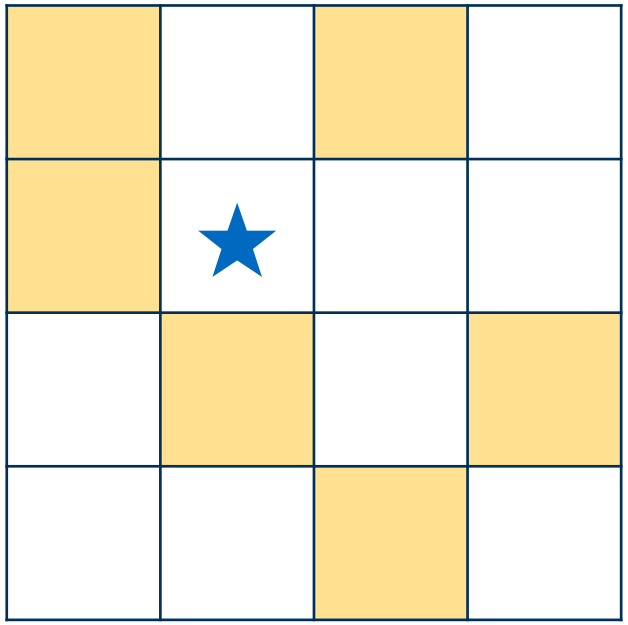
← Ensure non-negative visitation

- A policy is obtained from the occupancy measure $\pi(s,a) = \frac{x^*(s,a)}{\sum_{a'\in\mathcal{A}} x^*(s,a')}$.

Georgia Tech

Type 1: Diversionary Deception

- **Diversionary Deception:** A policy π_d is diversionary deceptive if the observer inference of the goal state is incorrect in no particular way.



★: True goal state
: The observer inference

Maximize accumulated reward

Get far from the original occupancy measure

$$\underset{x}{maximize} \quad \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \left[r(s,a)x(s,a) + \beta \left(x(s,a) - x^*(s,a) \right)^2 \right]$$

$$such\ that \quad \sum_{a\in\mathcal{A}} x(j,a) - \gamma \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \mathcal{T}(s,a,j)x(s,a) = \alpha(j)$$

$$\sum_{q\in\mathcal{S}_{goal}} \left[\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \mathcal{T}(s,a,q)x(s,a) + 1(q) \right] \geq v_{reach}$$

$$x(s,a) \geq 0, \qquad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$$

β is a weighting parameter between two objectives.

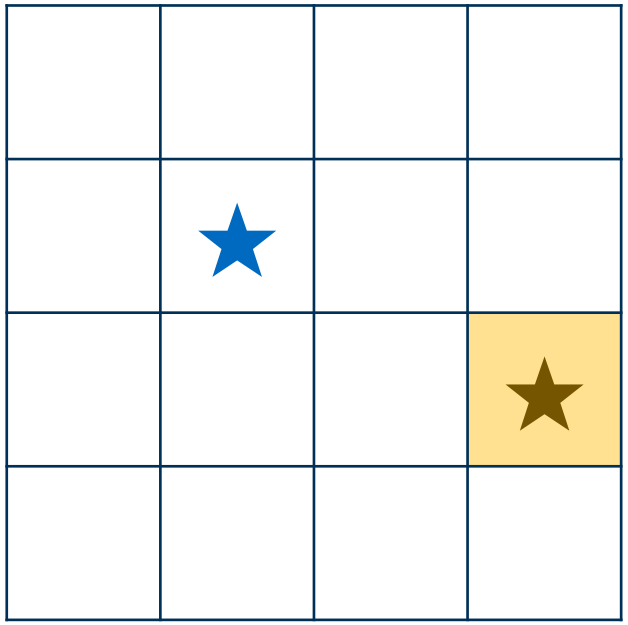
- Small β: Focus more on maximizing rewards. (Less deception)
- Large β: Focus more on being deceptive. (More deception)

Note: Linear programming becomes quadratic programming!

Georgia Tech

Type 2: Targeted Deception

- **Targeted Deception:** A policy π_d is targeted deceptive if the observer infers a user-specified incorrect state as the goal state.



★: True goal state
★: User-specified decoy state
: The observer inference

Maximize accumulated reward

Get closer to the target occupancy measure

$$\underset{x}{maximize} \quad \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[r(s,a)x(s,a) - \beta\left(x(s,a) - x_{tar}(s,a)\right)^2\right]$$

$$such\ that \quad \sum_{a\in\mathcal{A}} x(j,a) - \gamma\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\mathcal{T}(s,a,j)x(s,a) = \alpha(j)$$

$$\sum_{q\in\mathcal{S}_{goal}}\left[\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\mathcal{T}(s,a,q)x(s,a) + 1(q)\right] \geq v_{reach}$$

$$x(s,a) \geq 0, \qquad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$$

β is a weighting parameter between two objectives.

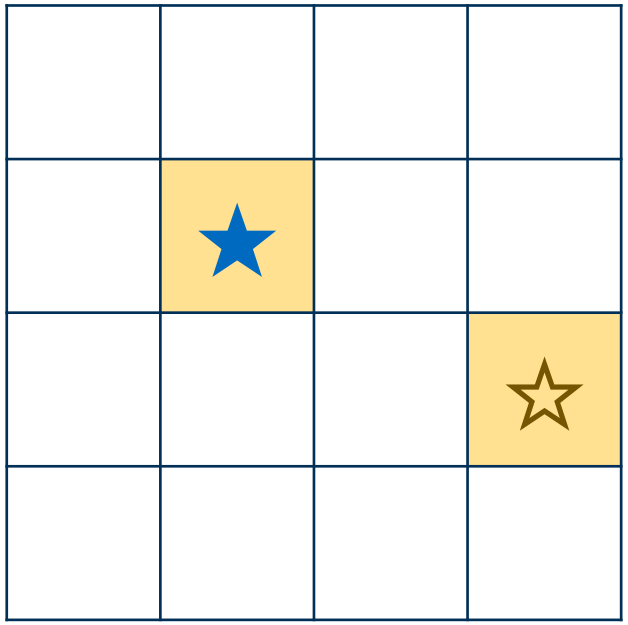
- Small β: Focus more on maximizing rewards. (Less deception)
- Large β: Focus more on being deceptive. (More deception)

Note: Linear programming becomes quadratic programming!

Georgia Tech

Type 3: Equivocal Deception

- **Equivocal Deception:** A policy π_d is equivocal deceptive if the observer infers that two potential goal states are equally likely.



★: True goal state
☆: User specified decoy state
▇: The observer inference

Maximize accumulated reward

Make the visitation equal in goal states and decoy goal states

$$\underset{x}{\text{maximize}} \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[r(s,a)x(s,a)\right] - \beta\left(\sum_{s\in\mathcal{S}_{goal}}\sum_{a\in\mathcal{A}} x(s,a) - \sum_{s\in\mathcal{S}_{decoy}}\sum_{a\in\mathcal{A}} x(s,a)\right)^2$$

$$\text{such that} \sum_{a\in\mathcal{A}} x(j,a) - \gamma\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\mathcal{T}(s,a,j)x(s,a) = \alpha(j)$$

$$\sum_{q\in\mathcal{S}_{goal}}\left[\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\mathcal{T}(s,a,q)x(s,a) + 1(q)\right] \geq v_{reach}$$

$$x(s,a) \geq 0, \qquad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$$

β is a weighting parameter between two objectives.

- Small β: Focus more on maximizing rewards. (Less deception)
- Large β: Focus more on being deceptive. (More deception)

Note: Linear programming becomes quadratic programming!

Question 1: How do we conceal the objective of the system under complete observation?

Georgia Tech

We examine the performance lost by deception

Question: Do we still have good performance while having deception?

- We define “revenue” as “total accumulated reward”.

$$R_\pi = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} r(s,a) x_\pi(s,a)$$

- When solving **diversionary deception**, loss in performance is bounded by:

$$\underbrace{R^* - R_\pi}_{\text{Loss in performance}} \leq \underbrace{\beta}_{\text{Level of deception}} \underbrace{\left(\sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} x^*(s,a)^2 + (1-\gamma)^2 \right)}_{\text{Constant with given MMDP}}$$

Note: This bound provides a design guideline for β, based on required performance.

- Bound for performance loss increases linearly as level of deception increase.
- This gives a “trade-off” between performance and deception.

Georgia Tech

We examine the performance lost by deception

- When solving **targeted deception**, loss in performance is bounded by:

$$\underbrace{R^* - R_\pi}_{\text{Loss in performance}} \leq \underbrace{\beta}_{\text{Level of deception}} \underbrace{\left(\sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} (x^*(s,a)^2 - 2x_{tar}(s,a)x^*(s,a)) - \frac{(1-\gamma)^{-2}}{|\mathcal{S}||\mathcal{A}|} + 2(1-\gamma)^{-1} \max_{s,a} x_{tar}(s,a) \right)}_{\text{Constant with given MMDP and user-defined } x_{tar}}$$

  - Bound for performance loss increases linearly with β. (Trade-off between performance and deception)
  - Choice of x_{tar} decides the constant coefficient of β.

- **Note: Design tip for x_{tar}**
  - x_{tar} should reflect the target objective to lead adversary to.
  - To reduce the coefficient of β, choose x_{tar} that is close to x^*.
  - To reduce the coefficient of β, keep the maximum value of x_{tar} small.

Note: This bound provides a design guideline for β, based on required performance.

Georgia Tech

We examine the performance lost by deception

- When solving **equivocal deception**, loss in performance is bounded by:

$$\underbrace{R^* - R_\pi}_{\text{Loss in performance}} \leq \underbrace{\beta}_{\text{Level of deception}} \underbrace{\left(\sum_{s \in \mathcal{S}_{goal}} \sum_{a \in \mathcal{A}} x^*(s,a) - \sum_{s \in \mathcal{S}_{decoy}} \sum_{a \in \mathcal{A}} x^*(s,a) \right)^2}_{\text{Constant with given MMDP}}$$

Note: This bound provides a design guideline for β, based on required performance.

  - Bound for performance loss increases linearly with β. (Trade-off between performance and deception)
  - The constant coefficient represents the difference of visitation in goal and decoy states.
    - When the visitation in goal and decoy states are similar under original, optimal strategy, equivocal deception is achieved with less performance loss.

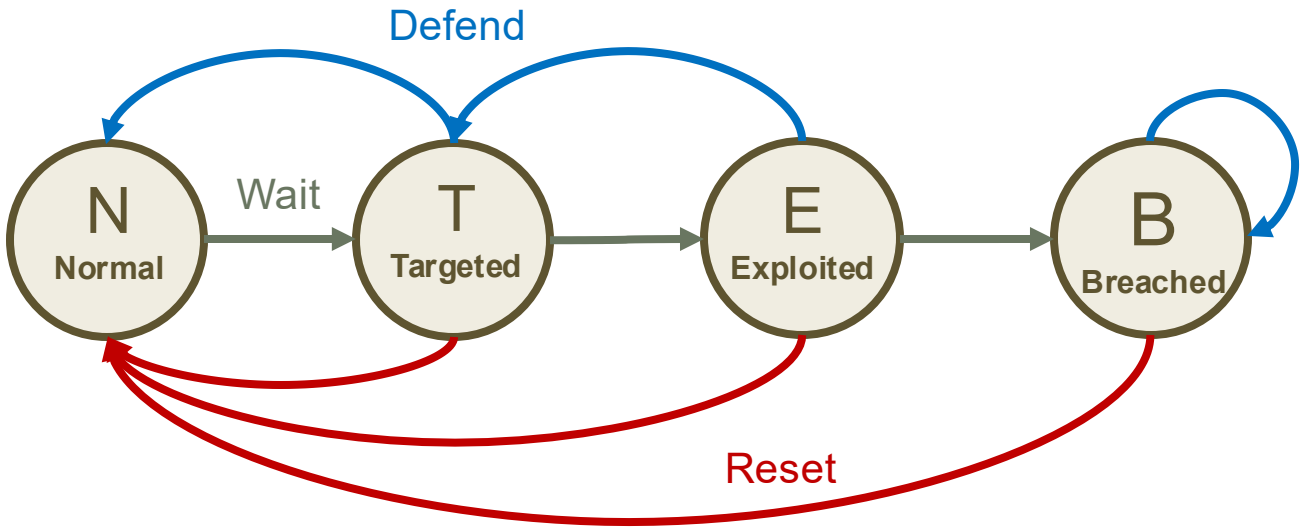
Question 2: Can we still achieve good performance while concealing the objective?

Georgia Tech

Application: Moving Target Defense

- Moving Target Defense (MTD)
  - A security technology to defend cyber attack by changing the configuration of a system
  - In **decoy-based MTD**, decoy systems are used to protect a real system by confusing attacker.

- Moving Target Defense can be formulated as an MDP



- Rewards of being in a state

$$Normal > Targeted > Exploited > Breached$$

- Rewards of taking an action

$$Wait > Defend > Reset$$

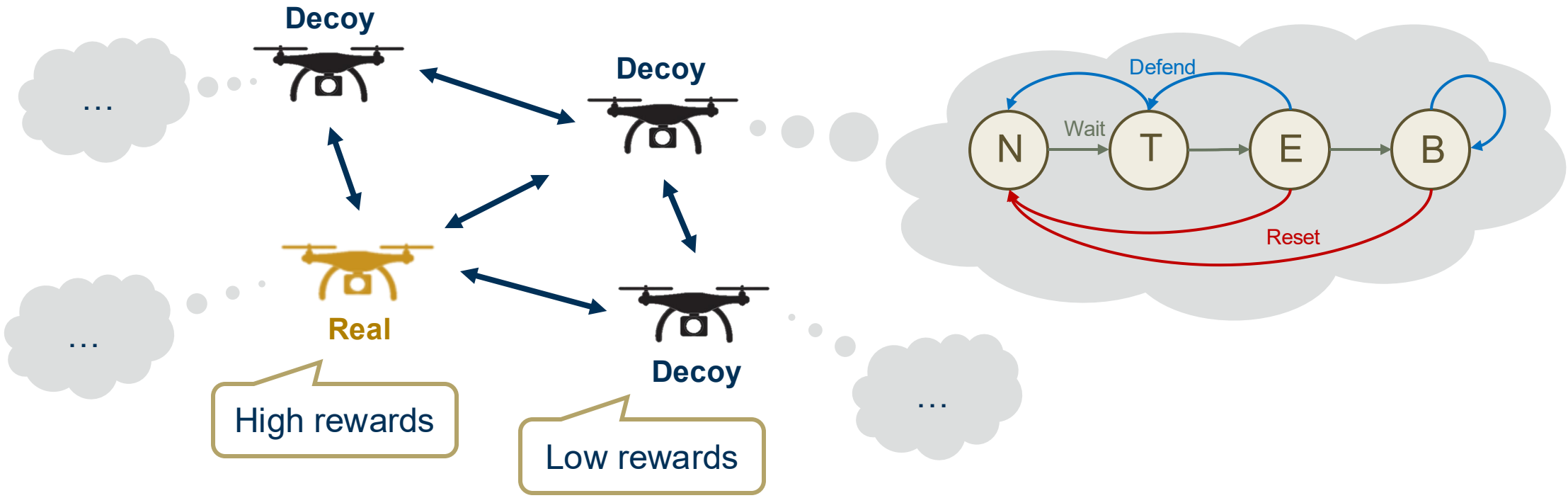
- **System objective (maximizing rewards)**
  - Be as normal as often as possible
  - Take less expensive actions

J. Zheng and A. S. Namin, "Markov decision process to enforce moving target defence policies," arXiv preprint arXiv:1905.09222, 2019.

Georgia Tech

Decoy system is used for confusion

- An example of decoy-based MTD consists of one real agent and three decoy agents. (No deception yet)
  - **System objective:** Protect real agent > Protect decoy agent



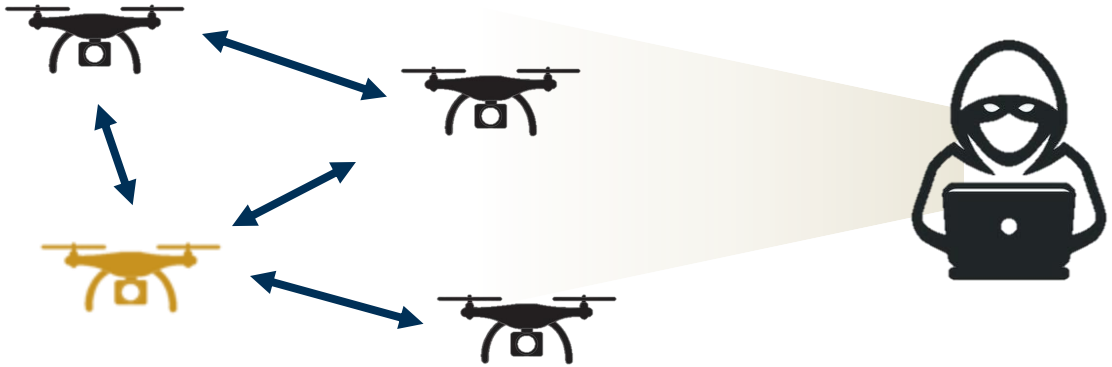
Goal: Deceive the adversary about the true target.

Georgia Tech

We applied deception to decoy-based MTD

Goal: Deceive the adversary about the true target.

- We applied deception to a decoy-based MTD



1. Decide the type of deception.
2. Solve the optimization problem to get deceptive policy π_d.
3. Take an action according to a deceptive policy π_d. (Add deception)
4. The adversary will observe the system and infer wrong rewards.

Question: How do we know that we deceived the adversary?

Georgia Tech

We defined metric to measure deception

Question: How do we know that we deceived the adversary?

- We deceived the adversary if its inference on the true target is wrong.
- Define relative importance of agents
  - Adversary's inference on the true target is based on the estimated rewards

$$D^i = \sum_{s^i \in \mathcal{S}^i} \omega(s^i)\tilde{R}(s^i), \qquad L^i = \frac{\exp(\frac{D^i}{\lambda})}{\sum_{j=1}^{M} \exp(\frac{D^j}{\lambda})}$$

Larger L^i : Agent i is more important

Smaller L^i : Agent i is less important

- From the adversary's perspective:



Observe the system to infer underlying rewards

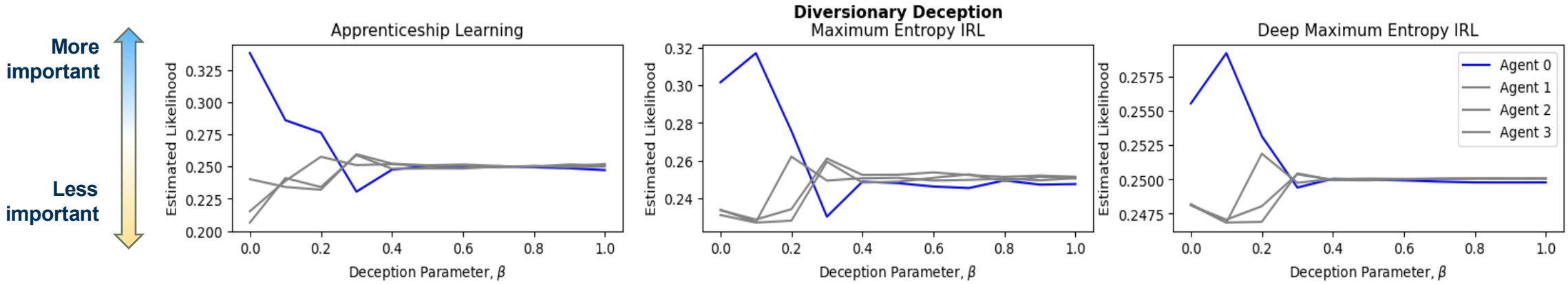
Compute relative importance of agents

Attack the most important agent

Georgia Tech

Effective deception can be achieved

- Deception is performed under three different adversarial models
- Diversionary deception
  - We want Agent 0 (real) to appear less important than the other agents.



- We become more deceptive when β gets larger.
- With large enough β, the adversary infers any decoy agent as the most important agent.

Takeaway: Agent 0 (real) becomes less important than other agents when $\beta \geq 0.3$.

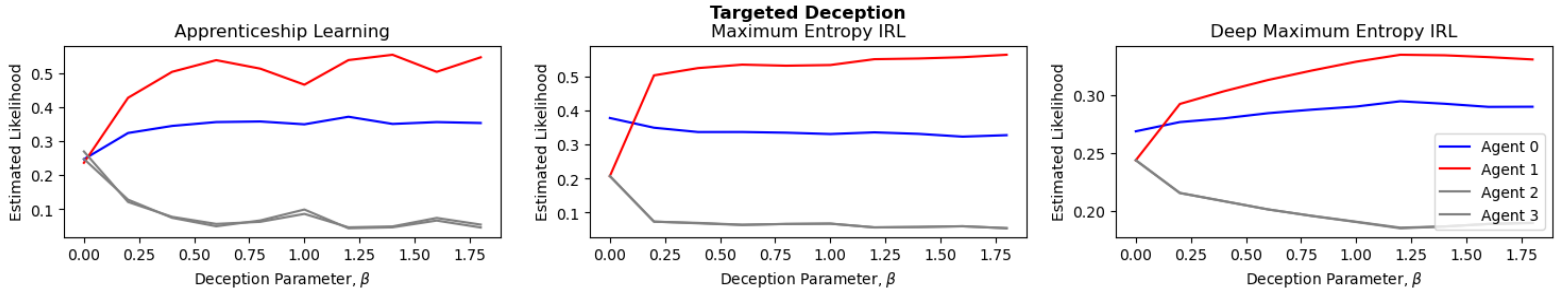
Georgia Tech

Effective deception can be achieved

- Targeted deception
  - We want the target decoy agent (Agent 1) to appear to be the most important agent.

More important

Less important

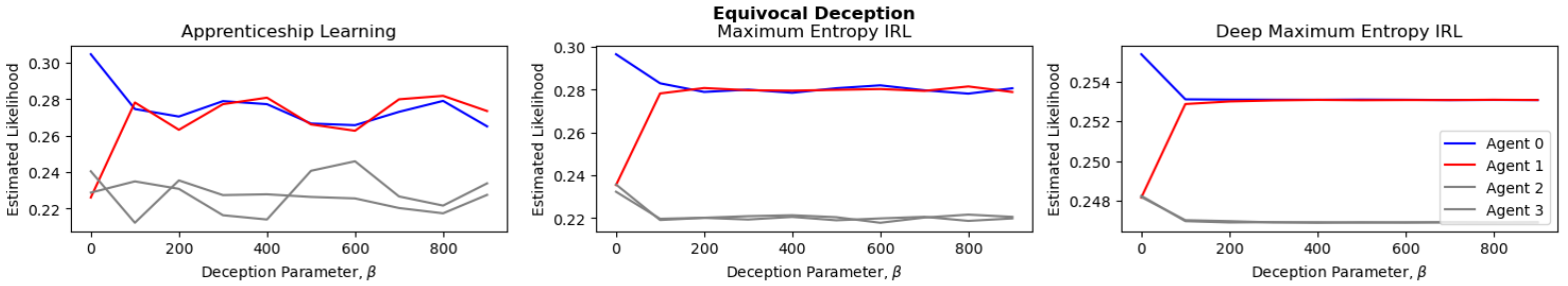


Takeaway: Agent 1 (decoy) becomes the most important agent when $\beta \geq 0.2$.

- Equivocal deception
  - We want the real agent (Agent 0) and the target decoy agent (Agent 1) to be equally important.

More important

Less important



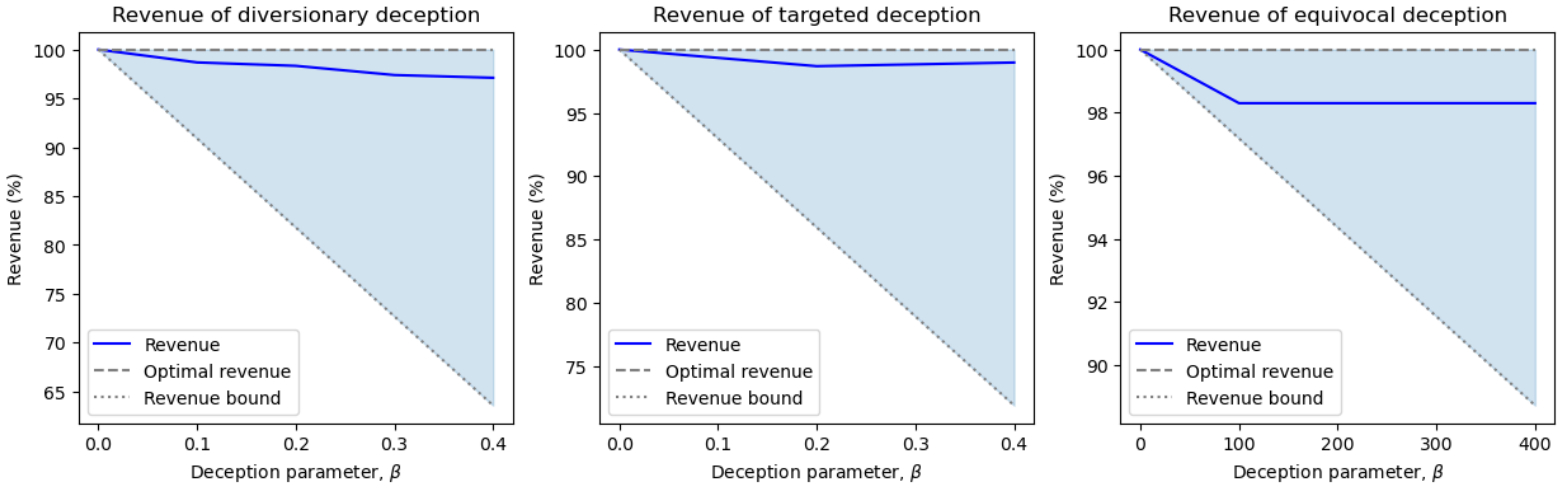
Takeaway: Agent 0 (real) and Agent 1 (decoy) become equally important when $\beta \geq 300$.

Question 1: How do we conceal the objective of the system under complete observation?

Georgia Tech

Good performance can be achieved with deception

- Total accumulated reward under deception



- In **diversionary deception**, the system achieves **97.39%** of accumulated reward when $\beta = 0.3$.
- In **targeted deception**, the system achieves **98.69%** of accumulated reward when $\beta = 0.2$.
- In **equivocal deception**, the system achieves **98.30%** of accumulated reward when $\beta = 300$.

Take away: Effective deception can be achieved with minimal loss in performance.

Question 2: Can we still achieve good performance while concealing the objective?

Thank you.

Email: yerinkim@gatech.edu

Georgia Tech