Security Analysis of Safe & Seldonian Reinforcement Learning Algorithms

A. Pinar Ozisik[1] and Philip S. Thomas[1]
University of Massachusetts[1]

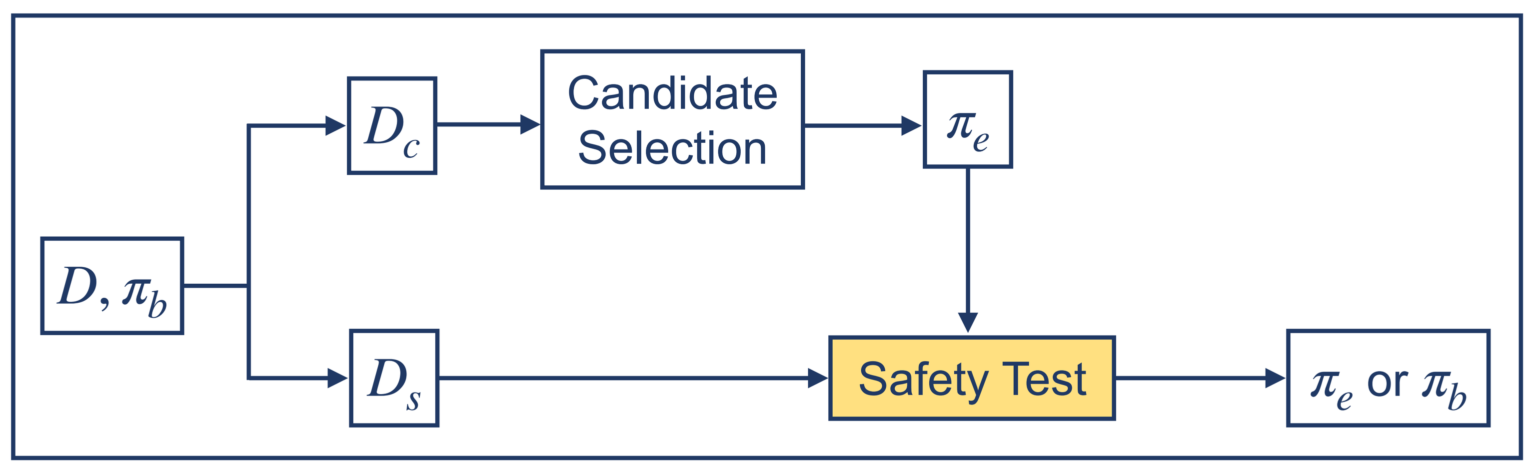
Problem Statement

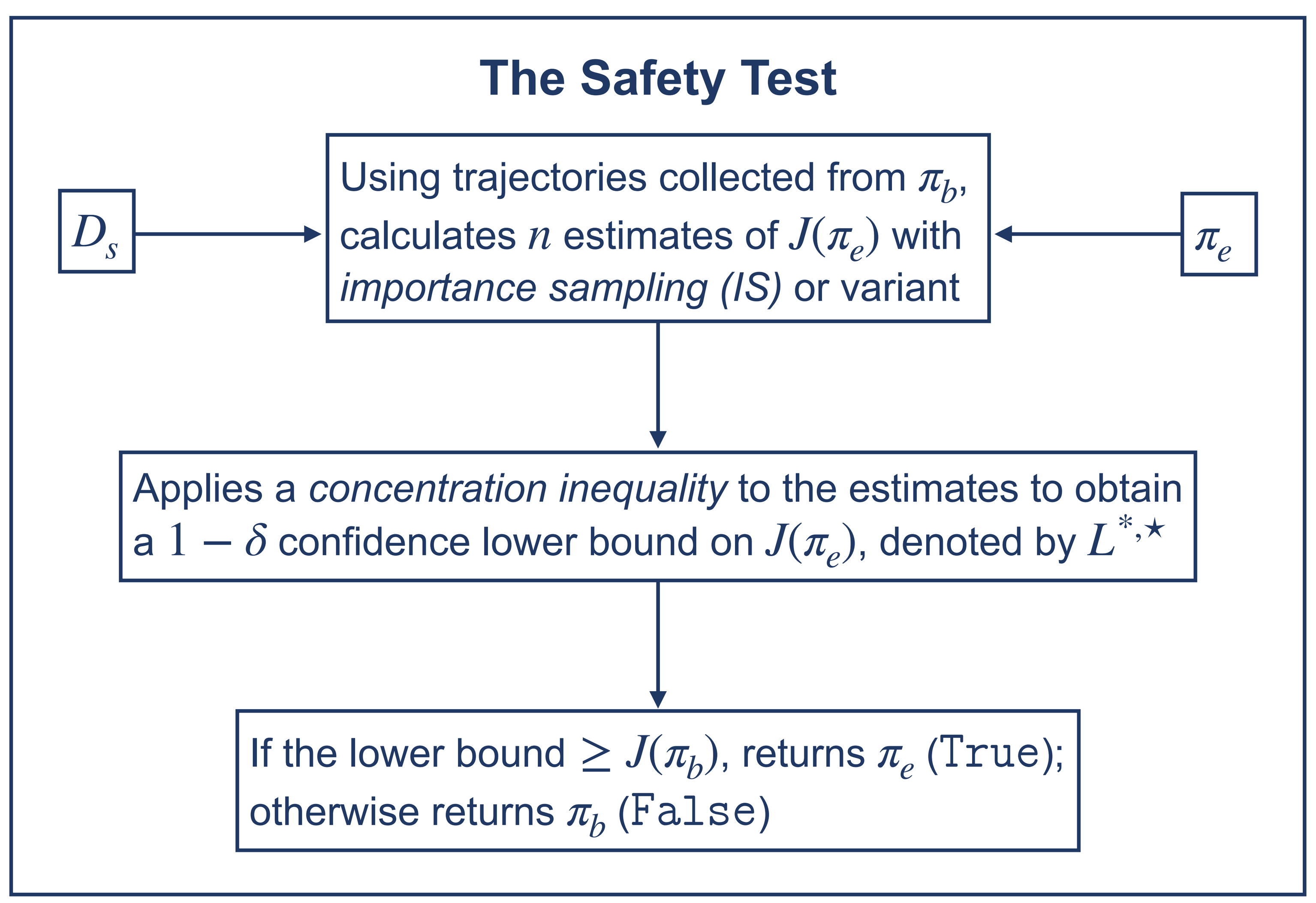
- RL is proposed for high-risk applications, such as improving type 1 diabetes and sepsis treatments
- Safe and/or Seldonian RL provides high-confidence guarantees that the application will not cause undesirable behavior
- **Limitation:** Assumes that training data is free from anomalies such as errors, missing entries, and malicious attacks
- **How robust are these algorithms to perturbations in data?**

Background

- **Goal:** Find a policy with larger performance than deployed policy π_b
- **Assumption:** Can collect data from policy π_b, but not from others
- Algorithm guarantees **safety:**

$$\Pr(J(a(D)) \geq J(\pi_b)) \geq 1-\delta$$





Problem Formulation

- Robustness to perturbations is analyzed by robustness to adversarial attacks
- Algorithms robust to adversarial attacks will also be robust to non-adversarial anomalies in data
- **Attacker model:** Attacker adds fabricated trajectories to dataset

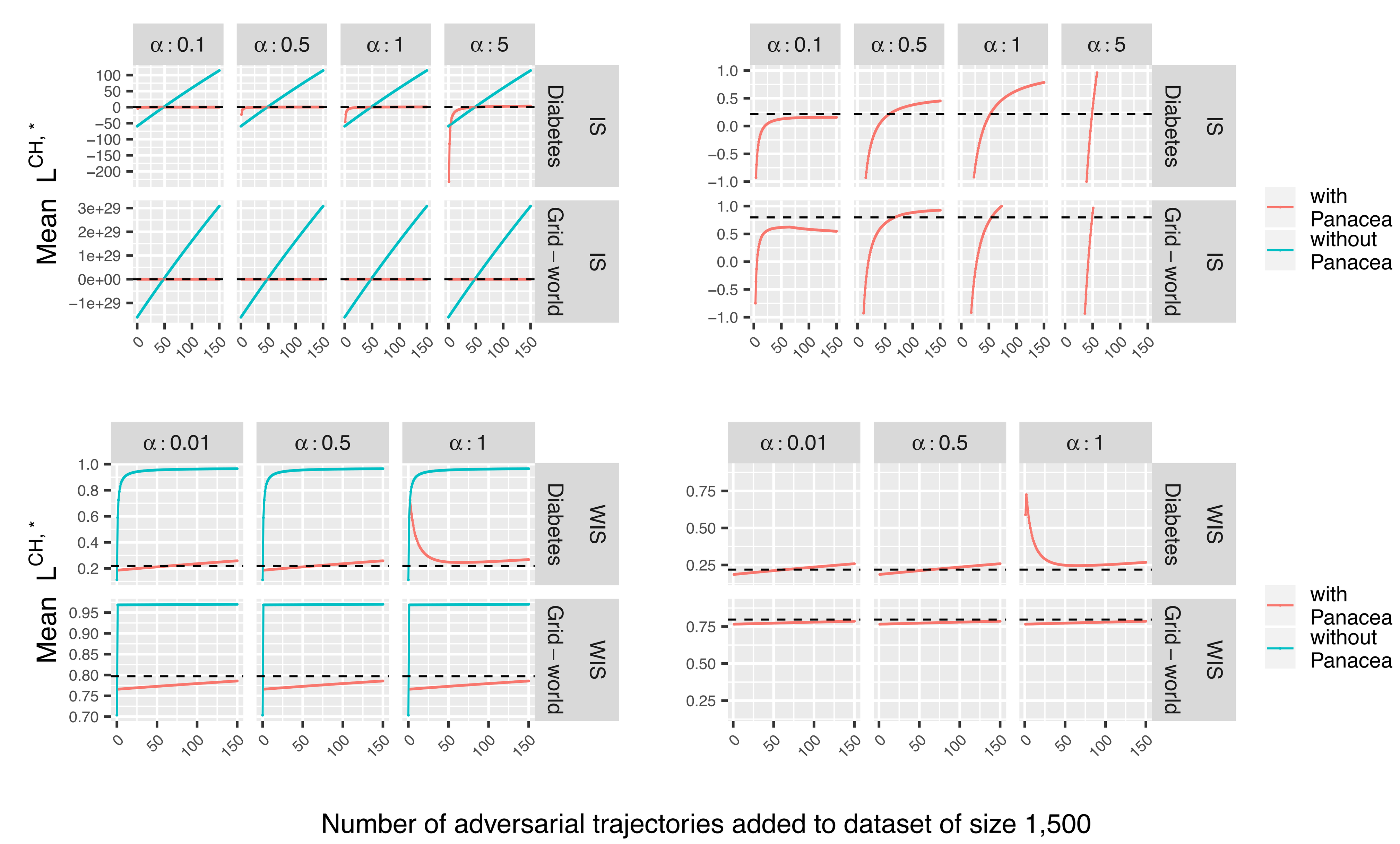
Panacea: A New Algorithm

- Named after the Greek goddess of universal health
- Provides α-security, with a user-specified α, if the number of corrupt trajectories in D is upper bounded
  - Takes as input number of corrupt trajectories
- Caps the importance weights using some clipping weight, c

α-security: A New Measure

- A measure that ensures safety even if data is corrupted by an adversary
- Larger α implies greater susceptibility to adversarial attacks
- **Assumption 1 (Inferior π_e).**
$$J(\pi_e) < J(\pi_b)$$
- **Assumption 2 (Absolute continuity).**
$$\big(\pi_b(s,a)=0\big) \implies \big(\pi_e(s,a)=0\big)$$
- **Assumption 3 (φ safety).** Given Assumption 1,
$$\Pr(\varphi(\pi_e, D, J(\pi_b)) = \texttt{True}) < \delta$$
- Under Assumptions 1, 2, and 3, a safety test, φ, is secure with constant α for π_e, π_b, k and D collected from π_b, where $|D|=n$, if and only if,
$$\forall m \in \mathscr{M}, \Pr\Big(\varphi\big(\pi_e, m(D,k), J(\pi_b)+\alpha\big) = \texttt{True}\Big) < \delta,$$
where m is an attack function and $\mathscr{M}$ is the set of all attack functions
- α is computed based on **best attacker strategy**
  - Causes maximum artificial increase in $1-\delta$ confidence lower bound on $J(\pi_e)$

Empirical Evaluation on Grid-world & Type 1 Diabetes Treatment



- Evaluated two safety tests
  - *Chernoff-Hoeffding* (CH) & IS
  - CH & *weighted importance sampling* (WIS)
- Randomly picked π_b and π_e
- **Point where blue line crosses the black-dotted line represents the number of trajectories needed to break existing safety tests**
  - It takes 49 trajectories for IS and 1 trajectory for WIS for both domains
- **Point where red line crosses the black-dotted line represents the number of trajectories needed to break Panacea**
  - *IS with $\alpha = 0.5$:* 59 trajectories in grid-world and 68 trajectories in diabetes domain
  - *WIS with $\alpha = 0.5$:* does not cross black-dotted line in grid-world and 65 trajectories in diabetes domain
- **Conclusion:** Safety tests can be extremely fragile, but Panacea provides user-specified robustness