CAREER: Robustifying Machine Learning for Cyber-Physical Systems

CNS-1845969, Mar 1, 2019 – Feb 29, 2024, PI: Soumik Sarkar, Iowa State University

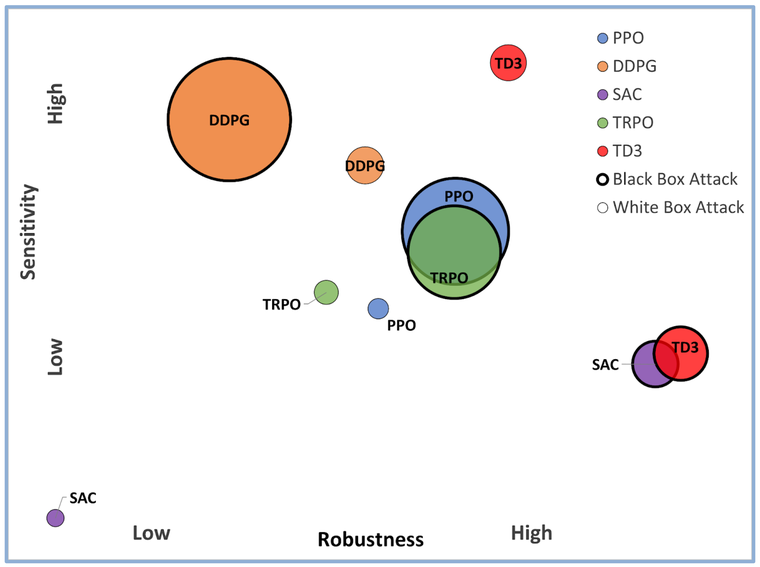
Challenge (Year 3 focus):

- Understand and quantify the robustness of deep reinforcement learning (DRL) approaches under various adversarial attacks and environments
- Develop machine learning (ML) based perception models with system-level safety awareness

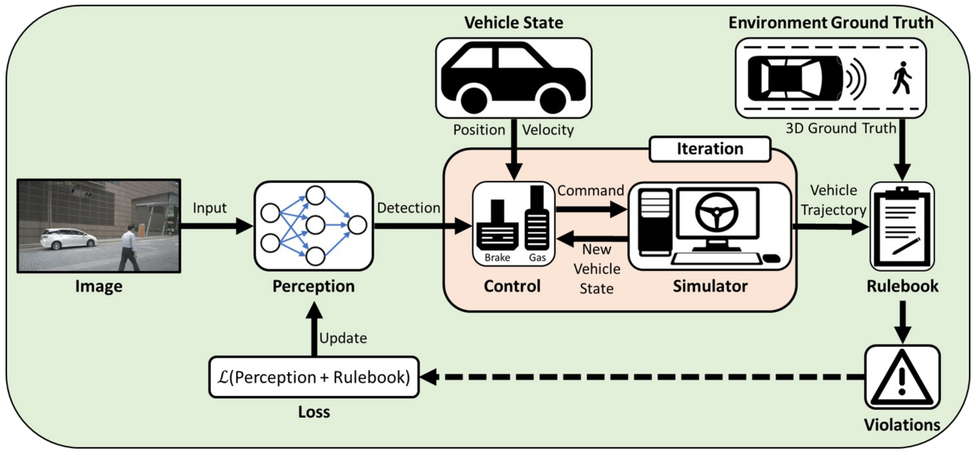
Solution:

- Empirical robustness study of various DRL approaches using additive perturbation based black box attacks and optimization based white box attacks on the action space and observation space
- Incorporate system-level safety objectives in ML model training via rulebook-based loss functions

Contact: soumiks@iastate.edu



Robustness study of DRL



Safety objectives for ML

Scientific Impact:

- Attack models studied are generic and applicable to any commonly used vision and RL-frameworks; Robust models can be deployed in various CPS applications that leverage ML modules
- System-level safety objectives in ML models can guarantee safety and improve CPS performance as a whole

Broader Impact:

- Motivation to develop and adopt more robust DRL algorithms in real life applications
- Bringing together ML and Formal methods is key to ensure safe use of ML modules in CPS operations
- Research results directly enriching a first of its kind CPS undergrad minor
- Partially supporting PhD study of 5 students including 3 URM students