Using AST to Identify Paths to Ethical Dilemmas

SISL
Stanford Intelligent Systems Laboratory

Penn Engineering

Idea

- Increasing use of safety-critical autonomous systems in society
- Autonomous agents might encounter decision situations where there's no clear ethical course of action
- Instead of trying to solve these dilemmas, we propose trying to avoid them altogether

Adaptive Stress Testing (AST)

Question: How does an agent arrive at a failure event?

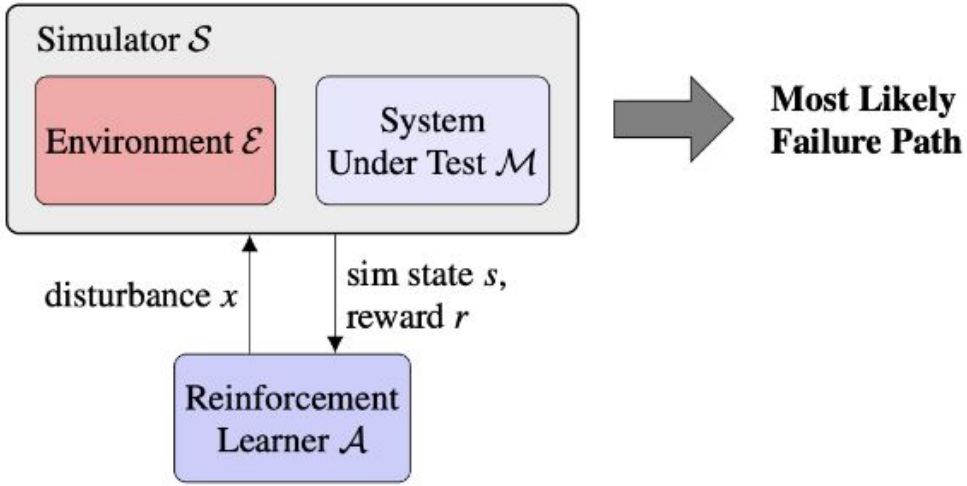

Figure 1: Simplified adaptive stress testing framework showing its core components (Lee et al. 2020).

Approach

1. **Define ethical dilemmas**
   a. Define set of ethical rules (e.g. "don't harm humans.")
   b. Action is unethical if it violates an ethical rule
   c. If all actions violate at least one ethical rule, agent finds himself in an ethical dilemma

2. **Applying AST**
   a. Failure event = ethical dilemma
   b. Find most likely paths to ethical dilemma in simulator

Toy Simulator Idea

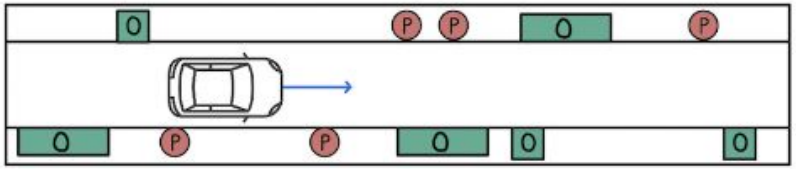

Figure 2: Example initial setup for simulator. The red circles depict pedestrians while the green boxes show immobile obstacles.

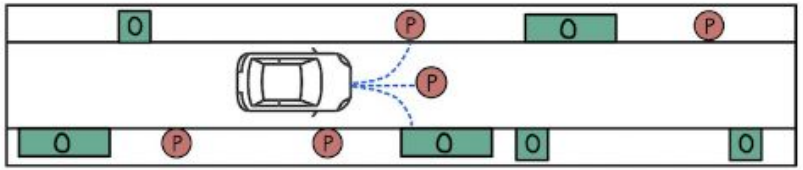

Figure 3: Example ethical dilemma. A pedestrian moves in front of the vehicle, leaving it with the option to crash into the pedestrian, a pedestrian on the left-hand side, or an obstacle on the right-hand side.

Future Research Directions

- Availability of simulators for ethical decision situations
- Context-specific, well-defined ethical rules necessary
- Downstream-effect of immediate decisions not taken into account in the current setup

Reach out if you have questions, feedback, …

akreuel@seas.upenn.edu, mark.c.koren21@gmail.com,
acorso@stanford.edu, mykel@stanford.edu

SISL
Stanford Intelligent Systems Laboratory

Penn Engineering