Indo/US Collaborative Research Grants

National Science Foundation of US and Technology Innovation Hubs of India

सत्यमेव जयते

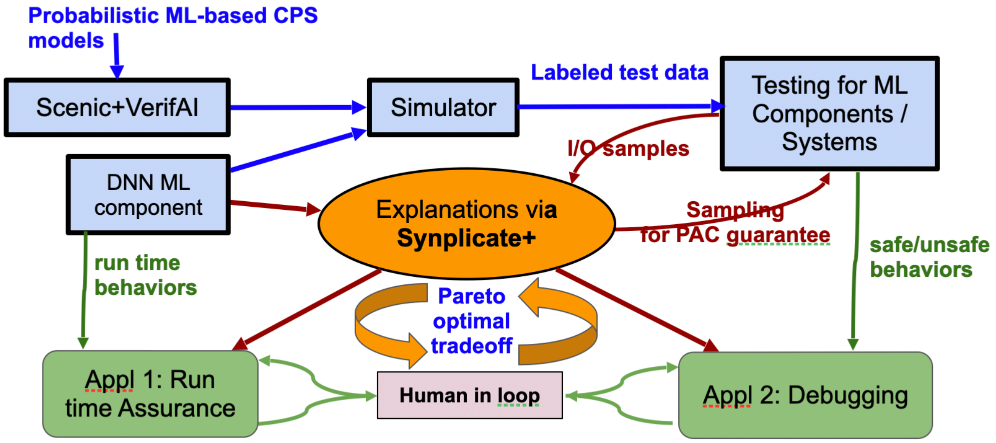
Title: Explaining Decision-Making in (Semi-)Autonomous Vehicles
Indian PIs: Prof. Supratik Chakraborty, Professor and PI, IIT Bombay, India.
Prof. Akshay S, Associate Professor and co-PI, IIT Bombay, India
US PI: Prof. Sanjit A. Seshia, University of California at Berkeley, CA, USA.

Autonomous and semi-autonomous vehicles are increasingly being used in many application domains. These systems strategically use machine learning (ML) components to perceive their environment and arrive at decisions that affect the health and safety of the vehicle. Unfortunately, very little is understood about the logic behind decision-making by these ML components. This is highly undesirable since an accident attributable to an erroneous decision taken by an ML component can have serious legal and ethical ramifications. It is, therefore, important to generate "explanations" of decisions taken by ML components, such that these explanations are credible and easily interpretable by humans.

The goal of this project is to provide an algorithmic, automatable, and parametrized framework for explanation generation for ML components used in autonomous and semi-autonomous vehicles. Since the behavior of ML components is often controlled by hundreds of thousands of parameters, finding human-understandable explanations in terms of these parameters is difficult. We, therefore, propose to view ML components as black boxes and introduce a systematic way to infer explanations from samples of their input-output behavior.

The project builds on the results of two other projects: ***VerifAI (UC Berkeley)*** *and* ***Synplicate (IIT Bombay, UC Berkeley).*** VerifAI, part of the **VeHiCal NSF-CPS frontier project,** develops techniques for the formal design and analysis of systems that include artificial intelligence (AI) and machine learning (ML) components. Synplicate is a framework for exploring human-understandable explanations of decisions taken by black-box ML models. Unlike earlier work, Synplicate provides end-users with tunability and a choice of tradeoff between understandability and accuracy, helping users explore the entire set of Pareto-optimal explanations.

Our project combines the techniques of VerifAI and Synplicate, augmenting them with new capabilities:
(i) debugging and improved understanding of the root cause of incorrect decisions taken by ML components in autonomous and semi-autonomous vehicles, (ii) facilitating runtime assurance with human-in-the-loop control in situations where an ML component is likely to take decisions that may lead to safety violations, (iii) investigating more scalable techniques for generating Pareto-optimal interpretations for a larger class of black-box models than it currently does. The overall flow of the project is shown below.


Probabilistic ML-based CPS models
Scenic+VerifAI
Simulator
Labeled test data
Testing for ML Components / Systems
I/O samples
DNN ML component
Explanations via Synplicate+
Sampling for PAC guarantee
run time behaviors
safe/unsafe behaviors
Pareto optimal tradeoff
Appl 1: Run time Assurance
Human in loop
Appl 2: Debugging