**Finding Safety Violations of AI-Enabled Control Systems through the Lens of Synthesized Proxy Programs**

JIEKE SHI, ZHOU YANG∗ , and JUNDA HE, Singapore Management University, Singapore BOWEN XU, North Carolina State University, United States DONGSUN KIM, Korea University, South Korea DONGGYUN HAN, Royal Holloway, University of London, United Kingdom DAVID LO, Singapore Management University, Singapore

**Aim of the Research Study**

This paper introduces Synthify, a novel framework aimed at identifying safety violations in AI-enabled control systems. Traditional testing of these systems is difficult due to their complex nature and unbalanced. The goal is to create a method that can efficiently find cases where the system’s behavior could lead to unsafe outcomes, especially in domains like autonomous vehicles or robotics where safety is critical.

**Methodology of the Study**

Synthify works by synthesizing proxy models that mimic the behavior of the original AI controller. These proxy models are trained using data collected from the AI-enabled system and are designed to approximate its decision logic closely. Once trained, they are used within a simulation environment to efficiently explore the system’s behavior across a wide range of input conditions. The approach combines data-driven learning with falsification search techniques to identify specific inputs that cause the system to violate safety properties. By decoupling the expensive original controller from the test process, Synthify speeds up the detection process without trading accuracy off. The authors evaluate their method using multiple benchmark control systems that integrate neural network controllers.

**Results of the Study**

The experimental results show that Synthify can find safety violations more efficiently and reliably than traditional falsification tools like Breach and S-TaLiRo. It achieves a higher success rate, requires fewer simulations, and is more effective at exploring the input space. These results suggest that proxy model synthesis significantly improves the ability to test AI-driven control systems without direct access to their internal logic.

**Implications for Research and Practice**

For researchers, this study demonstrates a practical method for tackling one of the most difficult challenges in AI system testing: black-box behavior. It opens up new horizons for integrating machine learning with formal verification. For practitioners, particularly in safety-critical industries, Synthify offers a scalable and effective tool for early detection of critical faults.