Toward Operational Safety Verification of AI-Enabled CPS (Student Abstract)

Imane Lamrani, Ayan Banerjee, Sandeep K.S Gupta
iMPACT Lab, CISDE
Arizona State University
Tempe, AZ
{ilamrani, abanerj3, sandeep.gupta}@asu.edu

Abstract

AI-enabled Cyber-physical systems (CPS) such as artificial pancreas (AP) or autonomous cars are using machine learning to make several critical decisions. The system is subject to inputs and scenarios which are not observed during training and the expected outputs are not known. Hence, popular model based verification techniques that characterize behavior of a control system before deployment using predictive models may be inaccurate and often result in incorrect safety analysis results. In addition, regulatory agencies are required to regulate safety-critical AI enabled CPS to ensure their operational safety. However, high complexity of the system result in myriad of safety concerns all of which may not only be comprehensively tested before deployment but also may not even be detected during design and testing phase. In this work, we propose a tool to help regulatory agencies compare the operation of the CPS with the specifications given by the manufacturer to ensure that the operation results conform with the safety assured design of a CPS.

Introduction

Recent cases of fatal failures of safety critical CPS have renewed the discussion on the certification problem. One important direction is the presence of artificial intelligence (AI) in the sub-components of the CPS. An AI agent that is deployed on a testing distribution that differs from the training distribution may not only exhibit poor performance, but also commit harmful or offensive actions. More broadly, AI enabled systems such as supervised classifiers can often suffer from the no oracle problem where the output for an unseen test case is non-deterministic and dependent on the environmental factors.Typically, the uncertainties or non-determinism in AI sub-components are not desirable in a safety critical component. As such, the coverage problem for AI enabled safety critical CPS can potentially encounter combinatorial explosion due to the presence of significant number of interacting external sub-components and environmental conditions of use cases. Verifying the safety and correct operation of these systems relies on verifying the correct interaction between the software and the physical environment (Leveson 2011). For example, dynamical variations between different and same individual of medical intelligent devices (e.g. AP control system) as well as the nonlinear nature of the dynamics of the physical system pose a major challenge in testing controllers of medical intelligent systems. Additionally, good environment models are often unavailable because of the high nonlinear variations present in the physical system due to different physiological conditions and operating conditions. Given the unsupervised nature of operation of intelligent systems, the operating conditions changes that are unaccounted for can guide towards misleading conclusions about the safety of these systems. On the other hand, industrial control systems are moving towards employment of advanced strategies such as adaptive control which uses feedback from the environment to update the control logic or the environmental model used by the controller to estimate the current state of the environment (Lamrani, Banerjee, and Gupta 2018a). Hence, additional and deeper safety analysis techniques must be developed as it is difficult for current safety verification methods to keep up with the increasing pace of technological change. In addition, safety critical CPS should meet government regulatory requirements before marketing. However, operational components interaction circumstances, inclusion of human-in-the-loop, and environmental changes results in myriad of safety concerns all of which may not only be comprehensibly tested before deployment but also may not even be detected during design and testing phase. In this paper, we refer

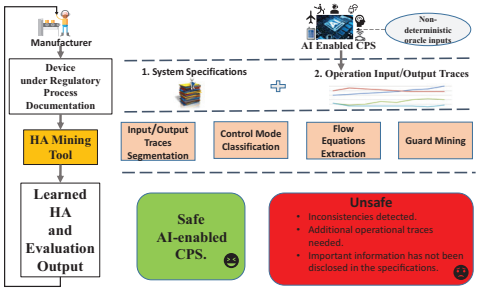

Figure 1: Overall Scheme of the Proposed Research.

to this problem as **operational safety verification problem**. For example, the Volkswagens defeat device that allowed vehicles to improperly meet US standards during regulatory testing (Tufekci 2015). In this paper, we propose a novel approach to solve the given problem of model based safety verification of AI enabled cyber-physical control systems with limited oracle. Our approach initially considers a hybrid system representation of the control system that describes the expected operation for which the system was tested, validated, and verified using controlled experimental studies. We then describe a methodology to periodically mine a hybrid system representation of the AI-enabled control system from input/output traces. The learned hybrid automata (HA) and the initial hybrid model defined in the documentation provided by the manufacturer are compared for safety verification purpose. The scheme of the proposed safety verification technique is shown in Figure 1. The HA mining algorithm takes the following inputs : 1) The time series traces obtained from the operation of the AI-enabled CPS, and 2) Documentation that contains general information including controller frequency, requirements, and design document. We use this documentation to model the initial HA of the AI-enabled system, if not provided in the system documentation. It employs a hybrid system mode segmentation methodology and density based clustering algorithm to derive the discrete mode transitions of the AI-enabled system. It employs Fisher information based analysis and Cramer Rao bound to derive the reset condition between two control modes. For each derived mode, it employs multi-variable polynomial regression analysis to derive the physical environment flow equations. The output of the HA mining algorithm is a learned non-linear HA. It then evaluate the consistency between the newly learned HA and the initial HA provided by the manufacturer. If learned HA is same as the initial hybrid model defined in the documentation provided by the manufacturer, then there is no change in the safety conclusion. However, if the learned hybrid system changes from the initially expressed one, then there might be a significant change in the safety conclusions.

Safety Verification HA-Mining

We propose a safety verification approach based on automated mining of hybrid automata from input/output traces collected from the operation of AI-enabled cyber-physical systems (Lamrani, Banerjee, and Gupta 2018b). An AI-enabled CPS is a system comprising a perception component, a planner/controller, and the environment (system under control) (Russell and Norvig 2016). Figure 2 shows the main steps of the proposed automated HA mining technique.

HA mining technique: The HA mining algorithm takes the observed continuous states of AI-enabled cyber-physical system inputs $\vec{x}$ and the control outputs $\vec{o}$ as inputs and extracts a hybrid system of the form of the tuple $< \mathcal{X}, \mathcal{M}, \mathcal{E}, \mathcal{I}, \mathcal{G}, \mathcal{R} >$ according to the definition of HA.

Periodic Mining Technique: The hybrid system mining is performed periodically. The learned hybrid system can differ from the specified system. The difference can be in the number of modes, flow dynamics, modes transitions, guard conditions, or reset conditions. There can be two reasons for this: a) the I/O traces do not allow derivation of characteristics of a given mode or dynamics, and b) the system received an unknown input for which the operation of the controller is uncertain. For the first case, we mine a partial hybrid system. However, parts that can be learned will be similar to the specified hybrid system and we can notify the manufacturer that additional traces are needed for the accomplishment of the safety verification process. In the second case, we will obtain a hybrid system that is different in its dynamics with the specified hybrid system. This entails that there is a new scenario that is being observed and the system has reacted in a way that is not expected using the specified hybrid system. In such a case, we consider utilizing the reach set analysis technique to derive the reach set and compare with the safety thresholds to re-evaluate the system safety in the near future (Alur et al. 1995). We have used the proposed approach in providing movement explanations for testing gesture based co-operative learning applications (Banerjee et al. 2019).

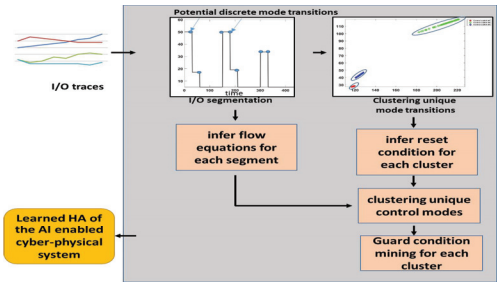

Figure 2: Hybrid Automata Mining Technique.

Evaluation and Results: In collaboration with Mayo clinic, we obtained continuous glucose monitoring readings and meal intake amounts from usage of Medtronic Minimed 670G. We simulate the artificial pancreas model using the UVA/Padova T1d platform to obtain the remaining inaccessible signals. We show the effectiveness of the proposed safety verification technique.

References

Alur, R.; Courcoubetis, C.; Halbwachs, N.; Henzinger, T. A.; Ho, P.-H.; Nicollin, X.; Olivero, A.; Sifakis, J.; and Yovine, S. 1995. The algorithmic analysis of hybrid systems. *Theoretical computer science* 138(1):3–34.

Banerjee, A.; Lamrani, I.; Paudyal, P.; and Gupta, S. 2019. Generation of movement explanations for testing gesture based co-operative learning applications. In *AITest'19*.

Lamrani, I.; Banerjee, A.; and Gupta, S. K. 2018a. Co-simulation of physical model and self-adaptive predictive controller using hybrid automata. In *STAF'18*. Springer.

Lamrani, I.; Banerjee, A.; and Gupta, S. K. 2018b. Hymn: Mining linear hybrid automata from input output traces of cyber-physical systems. In *ICPS'18*. IEEE.

Leveson, N. 2011. *Engineering a safer world: Systems thinking applied to safety*. MIT press.

Russell, S. J., and Norvig, P. 2016. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

Tufekci, Z. 2015. Volkswagen and the era of cheating software. *New York Times* 23.