Towards Building Reliable AI-Enabled Cyber-Physical Systems

Deyun Lyu[0000−0003−3017−7977]

Kyushu University, Fukuoka, Japan
lyu.deyun.107@s.kyushu-u.ac.jp

Abstract. AI-Enabled Cyber-Physical Systems (AI-enabled CPS), have recently demonstrated great application potential and become the subject of intense research. In such systems, AI controllers can monitor and control the behaviors of mechanical systems in real time, thanks to the "intelligence" of AI in handling various complex situations. Since such systems are usually deployed under safety-critical scenarios, building reliable AI-enabled CPS is of importance. Currently, this research faces many challenges, for example, the state-of-the-art testing methods and tools may be ineffective. Besides, difficulty in fault localization and lack of enhancement approach also pose great challenges to the quality assurance of such systems.
This paper summarizes our work in recent years, including a benchmark set, a falsification framework and some insights towards building reliable AI-enabled CPS. First, we constructed a benchmark set including 9 subject CPS collected from 7 industrial domains. Also, we conducted a comprehensive evaluation to the reliability and performance of these subject CPS. Then, we proposed a coverage-guided falsification framework `FalsifAI`, which fully utilizes 8 time-aware coverage criteria to guide the search of violation cases to the given specification. The experimental results demonstrate the effectiveness of `FalsifAI`. Finally, this paper gives some insights into building reliable AI-enabled CPS, aiming to raise more deep thinking and inspire works on this research.

Keywords: Software testing · Cyber-physical systems · AI controllers.

1 Introduction

AI-Enabled Cyber-Physical Systems (AI-enabled CPS) refer to the combinations of mechanical systems and AI controllers, in which AI controllers monitor and control the behaviors of mechanical systems in real time, according to the system states and external environments. Compared with classical CPS controllers, AI controllers are more "intelligent" when facing various complex situations. Certainly, AI controllers also have inherent drawbacks, such as their complexity and unaccountable decision logic. In practice, AI-enabled CPS are always deployed under safety-critical scenarios to perform complex control tasks, which raise the risk and probability of catastrophic accidents. Hence, quality assurance of AI-enabled CPS is of critical importance.

Currently, building reliable AI-enabled CPS faces many challenges. First, the state-of-the-art testing techniques and tools may be ineffective in finding violation cases to

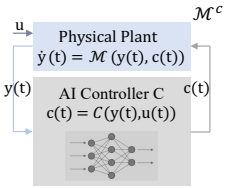

Fig. 1. AI-enabled CPS.

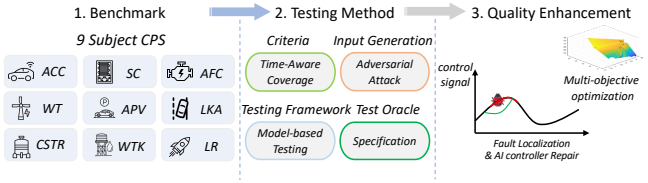

Fig. 2. Research route of building reliable AI-enabled CPS.

the given specifications. Falsification is a widely-applied testing method for quality assurance in CPS community, with the goal of seeking for an input signal that violates the given properties. In doing this, robustness provided by quantitative semantics of properties will be minimized until it turns negative, acting as a guidance of falsification. However, such guidance is mostly black-box, which makes the exploration blinded and blocks the understanding to the temporal internal behaviors of AI components. Second, how to diagnose and localize the faults of a violation case is also a burning issue. Specifically, identifying whether a control decision made by AI controller is correct or not remains to be studied, due to the lack of oracle for both system behaviors and control decisions. Besides, revealing the root causes for an erroneous decision is also a big challenge. Third, leveraging the diagnoses of violation cases to enhance the quality of AI-enabled CPS needs to be further explored.

Fig. 2 illustrates the flow of my research. Aiming at the above-mentioned problems, first, we created the first benchmark of AI-enabled CPS, including 9 subject CPS collected from 7 industrial fields, as well as systematic analysis to the dependability and performance of these AI-enabled CPS [4]. Then, we proposed a falsification framework `FalsifAI`, using 8 time-aware neuron coverage criteria as guidance [7]. The experimental result demonstrated the effectiveness of `FalsifAI`. Further, this paper offers some insights towards building reliable AI-enabled CPS.

2 An AI-Enabled CPS Benchmark

Considering the fact that there exist few benchmarks available for us to better solve the above-mentioned challenges, we created the first publicly accessible benchmark set with AI controllers trained by the state-of-the-art DRL algorithms [4]. The AI-enabled CPS in our benchmark set all follow the model in Fig. 1. This model $\mathcal{M}^c$ is composed of a physical plant and an AI controller $\mathcal{C}$. $\mathcal{M}^c$ takes input signal $\mathbf{u}$ and produces output signal $\mathcal{M}^c(u)$ if the whole system is viewed as a black box. The plant is a system whose dynamics are given by a black-box function $\mathcal{M}$, while AI controller $\mathcal{C}$ with the inputs $y(t)$ and $u(t)$ outputs a control command $c(t)$.

To obtain the overall understanding on the dependability and performance of AI-enabled CPS, we systematically evaluate these subject CPS including the evaluation and the falsification of some essential properties. We also explored the possibility of combining AI and classical controllers. Experimental results exhibited the great prospects of AI controllers, which also exposed some deficiencies of the current testing techniques.

3 A Coverage-Guided Falsification Framework

As mentioned in Sec. 1, classic falsification guided by robustness value is mostly black-box, which ignores the exploration of the inner behaviors of CPS with AI controllers. As an early attempt, we first studied the CPS with deep neural network (DNN) controllers and proposed a coverage-guided falsification framework `FalsifAI` with 8 time-aware coverage criteria.

3.1 Time-Aware Coverage Criteria

To capture the inner behaviors of DNN controllers and more thoroughly test such systems, we utilized and extended the concept *neuron coverage* [3]. If the output of a single neuron is greater than the set threshold, we say this neuron is covered. The more neurons are covered, the more diverse behaviors a DNN may act out. Notably, unlike the DNNs used as image classifiers, DNN controllers output time-series control decisions during simulation and operation. Hence, we considered the time domain features of DNN controllers and proposed 8 time-aware coverage criteria. Taking *Positive Differential Neuron Coverage (PDNC)* in our work as an example: if the activation value of a neuron increases by more than a threshold h within an interval I, we say this neuron is covered.

3.2 Two Loop Falsification Framework

The workflow of `FalsifAI` is composed of two loops, *exploration* and *exploitation*. The outer loop performs *exploration* by generating test cases which can increase our proposed coverage criteria, with an aim of exploring more different temporal behaviors. The inner loop focuses on *exploitation*, for the sake of finding violation cases near the test cases generated by exploration. This loop is based on classic falsification, which minimizes the robustness given by quantitative robust semantics. Fig. 3 illustrates the falsification success rate (out of 30) of `FalsifAI` on 6 CPS models, compared with two state-of-the-art falsification tools, namely Breach (Br) and S-Taliro (St), as well as a spatial-coverage guided approach Fal_Inp. `FalsifAI` significantly outperforms Br&St and has a performance close to or better than Fal_Inp.

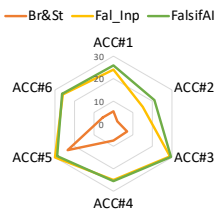

Fig. 3. Effectiveness of FalsifAI.

4 Insights on Building Reliable AI-enabled CPS

In this section, we briefly introduce current challenges not yet explored as well as our insights on how to solve them.

Fault localization. Fault localization plays an important role in CPS quality assurance. Bartocci et al. [1] proposed an approach to debug CPS models constructed by Simulink/Stateflow, by analyzing the given STL specification. As for AI-enabled CPS, this topic is still an untouched area. Mining useful patterns and hidden information from AI controller's states may be promising to tackle this problem.

Enhancement approach. For CPS with classical controller, engineers can analyze the modules with design flaws and redesign them according to control theory. However, unfortunately there is a lack of reliable and effective enhancement methods for AI-enabled CPS. Considering its data-driven design paradigm, neural network repair and ensemble learning are hopeful to enhance the reliability of such systems.

5 Related Work

In terms of *benchmark*, the annual competition ARCH-COMP [2] in CPS domain provides a series of CPS benchmarks. However, these benchmarks are either simple or do not contain AI components. There are also some work regarding *quality assurance* of AI-enabled CPS. [5] is a typical work of verification of such systems, which obtains reachable sets by reachability analysis. [6] attempts to falsify such systems using a gradient-based search method.

6 Conclusion and Future Work

This paper introduces our latest work about quality assurance of AI-enabled CPS, including a benchmark set, a coverage-guided falsification framework and some insights towards building reliable AI-enabled CPS. In the future, we will focus on fault localization and enhancement of such systems. We hope that this research can inspire more attempts to build reliable AI-enabled CPS. Moreover, we hope that industry applications directly related to this research, such as self-driving, will benefit from this work and in return, promotes the realization of *Industry 4.0*.

References

1. Bartocci, E., Ferrère, T., Manjunath, N., Ničković, D.: Localizing faults in simulink/stateflow models with stl. In: HSCC'18. pp. 197–206 (2018)
2. Ernst, G., Arcaini, P., Bennani, I., Chandratre, A., Donzé, A., Fainekos, G., Frehse, G., Gaaloul, K., Inoue, J., Khandait, T., et al.: Arch-comp 2021 category report: Falsification with validation of results. In: ARCH@ ADHS. pp. 133–152 (2021)
3. Pei, K., Cao, Y., Yang, J., Jana, S.: Deepxplore: Automated whitebox testing of deep learning systems. In: SOSP'17. pp. 1–18 (2017)
4. Song, J., Lyu, D., Zhang, Z., Wang, Z., Zhang, T., Ma, L.: When cyber-physical systems meet ai: A benchmark, an evaluation, and a way forward. In: ICSE'22-SEIP. pp. 343–352 (2022)
5. Tran, H.D., Yang, X., Manzanas Lopez, D., Musau, P., Nguyen, L.V., Xiang, W., Bak, S., Johnson, T.T.: Nnv: the neural network verification tool for deep neural networks and learning-enabled cyber-physical systems. In: CAV'20. pp. 3–17. Springer (2020)
6. Yaghoubi, S., Fainekos, G.: Gray-box adversarial testing for control systems with machine learning components. In: HSCC'19. pp. 179–184 (2019)
7. Zhang, Z., Lyu, D., Arcaini, P., Ma, L., Hasuo, I., Zhao, J.: Falsifai: Falsification of ai-enabled hybrid control systems guided by time-aware coverage criteria. TSE'22 (2022)