On the Safety of Automotive Systems Incorporating Machine Learning based Components A position paper

Mohamad Gharib*, Paolo Lollini*, Marco Botta†, Elvio Amparore†, Susanna Donatelli†, Andrea Bondavalli*
* University of Florence, Italy
{mohamad.gharib,paolo.lollini,andrea.bondavalli}@unifi.it
† University of Torino, Italy
{botta,amparore,susi}@di.unito.it

***Abstract*—Machine learning (ML) components are increasingly adopted in many automated systems. Their ability to learn and work with novel input/incomplete knowledge and their generalization capabilities make them highly desirable solutions for complex problems. This has motivated the inclusion of ML techniques/components in products for many industrial domains including automotive systems. Such systems are safety-critical systems since their failure may cause death or injury to humans. Therefore, their safety must be ensured before they are used in their operational environment. However, existing safety standards and Verification and Validation (V&V) techniques do not properly address the special characteristics of ML-based components such as non-determinism, non-transparency, instability. This position paper presents the authors' view on the safety of automotive systems incorporating ML-based components, and it is intended to motivate and sketch a research agenda for extending a safety standard, namely ISO 26262, to address challenges posed by incorporating ML-based components in automotive systems.**

***Index Terms*—Automotive systems, Functional safety, Machine learning, ISO 26262, ADAS**

I. Introduction

The last few years have witnessed an increasing adoption of Machine learning (ML) components in many automated systems covering almost all the main domains of our lives [1]. Their ability to learn and work with incomplete knowledge [2], and their generalization capabilities make them highly desirable solutions for complex problems [3]. This has motivated many system manufacturers to adopt ML-based components in their products, and the use of such components has emerged in many industrial domains (e.g., medical, automotive), performing complex tasks such as pattern recognition, image recognition, and even control [3]. However, most of the systems in these domains are classified as safety-critical systems, where their failure may cause death or injury to humans [4]. Thus, their safety should be assessed and assured before they are used in their operational environment [5].

Typically, the safety of such systems can be assessed by their compliance with safety standards [5], [6], such as the Road vehicles - functional safety (ISO 26262) [7]. Yet the adequacy of such standards for ML remains controversial since they, usually, do not address the special characteristics of ML-based components such as the presence of non-determinism, limited or missing transparency, uncertain error rate, instability of results [3], [8], [9]. Not only the certification of ML for safety-critical systems is problematic, but also their Verification and Validation (V&V), mostly due to the same previously mentioned ML special characteristics [3], [10], [11].

In particular, the traditional concept of safe software and V&V process does not apply to machine learning. The fact that the software is correct (it is free from bugs) does not imply that the recognition process will behave as expected since it can fail due to several reasons including the limitations of the machine learner itself. For example, a recent article showed that the change of a single pixel on the frame of a camera could lead the machine learner to completely different decisions [12], without exploiting software bugs but only because of the nature of the trained network.

Currently, the automotive industry is reacting to this problem with the development of a new standard, SOTIF (Safety of the intended functionality) [13], which is currently in its draft stages. However, this standard is still far from proposing a definitive approach to the safe use of machine learning. Other domains, where the application of machine learning and autonomy is still limited, are much behind. *As per today, there is neither a safety standard for certifying automotive systems that incorporate ML-based components, nor a concrete agreed upon method for their V&V*. Thus, there is no sufficient evidence to assure their safety.

This position paper presents the authors' view on the safety of automotive systems incorporating ML-based components. We discuss the limitations of a well-adopted safety standard, namely ISO 26262/SOTIF[1], and then the activities required to extend it to address challenges posed by incorporating ML-based components in automotive systems. The rest of the paper is organized as follows; Section II presents the research baseline, and we describe an illustrative example in Section III. In Section IV, we present the problem statement and research questions, while we discuss the possible solutions in Section V. Finally, we conclude the paper in Section VI.

[1] The focus will be on ISO 26262 since SOTIF is currently in its draft stages and not publicly available

II. Research Baseline

A. ISO 26262

ISO 26262 [7] is a functional safety standard developed with the main objective to provide guidelines and best practices to increase the safety of Electronic and Electric (E/E) systems in vehicles. It covers the overall automotive safety life cycle, and it focuses on the identification of hazards related to E/E systems and their associated risks. Then, the associated risks are assigned an Automotive Safety Integrity Level (ASIL), which can be classified under, Quality Management (QM[2]), ASIL A, ASIL B, ASIL C, and ASIL D, where ASIL D requires the highest risk reduction effort.

As previously mentioned, the adequacy of ISO 26262 for ML remains controversial since it, usually, does not address the special characteristics of ML-based components [1]. More specifically, ISO 26262 deals with ML-based components as a traditional software system (Part 6 of the ISO 26262 - product development at the software level). Despite this, some attempts have been made to provide recommendations/suggestions on how to adapt some of these standards to accommodate ML (e.g., [1], [11]).

B. Machine Learning and Safety

Recent advances in ML algorithms, like Deep Learning (DL), have resulted in an impressive performance boost for a broad spectrum of complex, real-world tasks including object recognition, image segmentation, and speech recognition. ML-based components are increasingly getting deployed in safety-critical systems and applications such as self-driving cars [14], automated passenger screening, and medical diagnosis. Artificial Neural Networks (ANN) are inspired by the physical structure of the human brain [15], and they are the basis of modern DL systems. ANN are organized into well-defined structures, such as discrete layers, grids, or controlled loops.

Engineering a safe system always carries some level of epistemic uncertainty, once it is dependent on the knowledge that the people involved in the safety assessment have of all the different aspects of the system, including its expected behavior. With the introduction of ML, the challenge of managing epistemic uncertainty is increased by the impossibility of being certain that the ML algorithm will make a correct prediction. An ML algorithm makes predictions based on a model calculated from its input data. Inherently, these predictions carry some chance of being wrong. Learning is an inductive process; whereas in deduction one can guarantee that a conclusion follows from its premises; the same does not apply in induction. Making the case that a system whose behavior cannot be fully predictable is sufficiently safe is not obvious. It breaks the classical paradigm of high-integrity systems' development. The natural step forward is then to understand the factors contributing to the accuracy of ML systems and if and how it is possible to limit the magnitude of misbehavior.

The use of learned components in safety-critical systems poses several challenges. On one hand, the complexity of deep neural networks, with millions of parameters, makes it difficult to verify and test such systems. Also, the traceability of safety requirements is lost in the network. At the same time, their ability to generalize to novel situations is useful in many tasks, but a major challenge during verification and validation. Although there are several safety principles and methods for engineering safer systems in the literature (e.g., [9], [16], [17]), no proper effort has been devoted on how they should be adapted to address challenges posed by the use of ML-based components.

III. Illustrative Example: Maneuver Assistance System

Our illustrative example is an Advanced Driver Assistance Systems (ADAS), namely Maneuver Assistance System (MAS) that is expected to increase the drivers safety by detecting and preventing unintended maneuvers[3].

MAS will collect information about the vehicle, its surroundings, and the driver's 1- head pose and motion that is used to identify the driver's visual orientation and predict some of its maneuvers (e.g., head motion may precede a maneuver); 2- hands and foot location and motions that is used to predict some driver's actions. This information is analyzed by an ML-based component to identify whether the maneuver is intended or not depending on an already learned *intended*/*unintended* maneuvers patterns, i.e., if there is a need and/or desire for such maneuver it is considered an *intended* one. Otherwise, it is considered as an *unintended* one. Accordingly, MAS should allow *intended* maneuvers and prevents *unintended* ones depending on lock actuator.

MAS is a good example to represent an automotive system that incorporates ML-based component, where its failure to identify *intended*/*unintended* maneuvers and act accordingly in a reasonable time can lead to life-threatening situations.

IV. Problem Statement and Research Questions

Considering the previous example, the ML-based component may wrongly categorize an intended maneuver as an unintended one, which will prevent the driver from performing an intended maneuver, or it may wrongly categorize an unintended maneuver as an intended one, which will allow an unintended maneuver to be performed. Both of these situations can be life-threatening. To mitigate such threat, MAS should be able to minimize the wrong identification of the type of the maneuver and it should also be able to minimize the consequences of wrong identification when occurred.

In general, we need to research and investigate how ML-based components can be used safely in automotive systems. More specifically, any automotive system that incorporates ML-based components should be designed in a way that minimizes the ML-based component related errors, faults and failures. Moreover, it should be able to assess the occurrence

[2] QM does not require a risk reduction effort

[3] For more information about the example, please refer to [18]

and consequences of such errors, faults and failures. Finally, it should be able to mitigate any safety-related consequences due to such errors, faults and failures. In order to achieve that, we need to answer the following questions:

- *RQ1: How ML-based components are different from traditional software components with respect to safety-related aspects?* As previously mentioned, although if an ML-based component is free from bugs it still can fail to behave correctly due to several reasons.
- *RQ2: How the main safety-related concepts should be modified to find an adequate context to limit the safety risks of automotive systems that incorporate ML-based components?* Since ML-based components and standard software components are different in nature, does existing safety-related concepts (e.g., errors, faults, failures) applies to ML-based components as they are, or they need to be adapted and modified to capture the challenges posed by the use of ML-based components?
- *RQ3: How can we assess the likelihood and severity of ML-based components errors, faults and failures?* As any safety-related consequences of ML-based components errors, faults and failures should be mitigated, we need to provide techniques enabling such assessment.
- *RQ4: Do existing safety principles applies to the design of automotive systems that incorporate ML-based components?* Several safety principles for engineering safer systems have been proposed[4]: do such principles apply to assure the safe design of automotive systems that incorporate ML-based components as they are or they need to be adapted?
- *RQ5: How the ISO 26262 standard should be extended to address the use of ML-based components?* As previously mentioned, ISO 26262 deals with ML-based components as a traditional software system. Therefore, we need to investigate how it should be extended to address challenges posed by incorporating ML-based components in automotive systems.

V. Proposed Solutions and Discussion

In this section, we present and discuss our research agenda along with our ideas to answer the research questions. The research agenda has been structured in a process of five steps (depicted in Fig. 1), each of these steps aims to answer its corresponding research questions:

- *Step 1. Investigating the specific characteristics of ML-based components with respect to safety-related aspects:* ML makes predictions based on a model calculated from its input data. Inherently, these predictions carry some chance of being wrong. Learning is an inductive process; whereas in deduction one can guarantee that a conclusion follows from its premises, the same does not apply in induction. Therefore, it is required to understand the factors that might contribute to the errors of ML predictions and if and how it is possible to limit such errors. In particular, the specific characteristics of ML to be investigated are their non-determinism (i.e., our inability to predict which is the output for a specific input), non-transparency (i.e., our inability to scrutinize the reasons of a given decision), their error rate (i.e., we can only estimate the final error rate with a given degree of confidence), and their instability (i.e. when small change in the input, typically not significant for the human, produces very different output).
- *Step 2. Developing a conceptual model for modeling automotive systems incorporating ML-based components:* in order to design safer systems, first we should be able to capture/model their main safety aspects. Therefore, the results of the analysis performed at step 1 will be used to develop a conceptual model that identifies the main safety concepts (e.g., errors, faults, failures) of both automotive systems and ML-based components, the different interrelations among these concepts, and how such concepts should be adapted/modified to be used for modeling automotive systems incorporating ML-based components. One main purpose of this model is reducing the cognitive complexity while incorporating ML-based components in automotive systems by providing simple but expressive concepts for modeling the main safety-related aspects. In addition, a Model Driven Engineering (MDE) framework, and a supporting tool could be developed at this step to facilitate the design of automotive systems incorporating ML-based components.
- *Step 3. Developing safety assessment techniques:* the conceptual model developed in step 2 will be used as a base to develop techniques for assessing the safety of automotive systems incorporating ML-based components. These techniques will enable to detect the occurrence of ML errors, faults and failures as well as assessing their consequences, which can be used for avoiding, tolerating and/or mitigating any safety-related consequences that may arise due to such errors, faults and failures. We believe developing model-based stochastic techniques will be a good option for enriching the MDE to perform the required safety assessment activities.
- *Step 4. Developing safety principles for assuring the safe incorporation of ML-based components in automotive systems:* taking into consideration the safety assessment techniques developed in step 3, we will conduct an extensive analysis of existing safety engineering principles to identify which of such principles can be adopted as they are, when and how they need to be adapted to address the characteristics of ML-based components, and when novel safety principles specialized for automotive systems incorporating ML-based components need to be developed. The final result of this activity will be a set of safety principles specialized for assuring the safe incorporation of ML-based components in automotive systems. This list will be used to develop architectural principles and design guidelines for assuring the safety of automotive systems incorporating ML-based components (e.g., fail-

[4] For extensive list of safety principles please refer to [17]

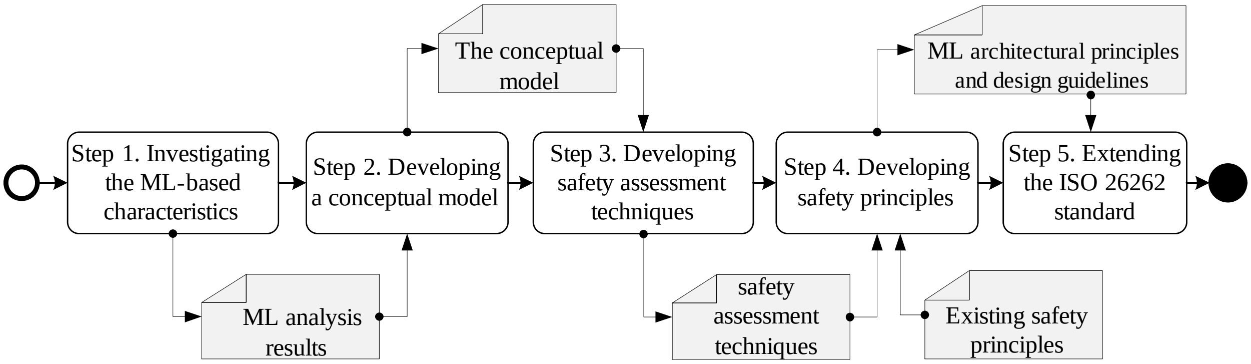
The conceptual model
ML architectural principles and design guidelines
Step 1. Investigating the ML-based characteristics
Step 2. Developing a conceptual model
Step 3. Developing safety assessment techniques
Step 4. Developing safety principles
Step 5. Extending the ISO 26262 standard
ML analysis results
safety assessment techniques
Existing safety principles

Fig. 1. The research process

safe depending on redundancy and/or segregation).

- *Step 5. Extending ISO 26262 standard to address the use of ML-based components:* we will investigate how the safety standard ISO 26262 should be adapted/extended to consider the proposed solutions developed within this research. In particular, the authors will mainly focus on Part 6 of the ISO 26262 - product development at the software level, trying to extend and modify its Clauses (6-6, 6-7, 6-8, 6-9, 6-10, and 6-11) in accordance to the findings and results of this research. In addition, we intend to consult or even involve members of the ISO 26262 standardization body in this step.

VI. Conclusion

In this position paper, we argued that the safe use of ML-based components in automotive systems must be assessed and assured before as such systems are classified as safety-critical systems, since ML-based components are subject to uncertainty and instability due to the nature of the trained network as well as their execution environment. Our argument is based on the ML and safety related literature concerning the automotive domain in which, we were not able to find neither a safety standard for certifying automotive systems that incorporate ML-based components, nor a concrete agreed upon method for their V&V. In addition, we presented our view on the safety of automotive systems incorporating ML-based components. Then, we have motivated and provided a research agenda for extending the ISO 26262 (Road vehicles - functional safety standard), in order to address challenges posed by incorporating ML-based components in automotive systems.

Acknowledgment

This work has been partially supported by the “Ente Cassa Di Risparmio di Firenze”, Bando per progetti 2016, and by the REGIONE TOSCANA POR FESR 2014-2020 SISTER “SIgnaling & Sensing Technologies in Railway application”.

References

[1] R. Salay, R. Queiroz, and K. Czarnecki, “An Analysis of ISO 26262: Using Machine Learning Safely in Automotive Software,” *arXiv preprint*, sep 2017.

[2] Z. Kurd, T. Kelly, and J. Austin, “Developing artificial neural networks for safety critical systems,” *Neural Computing and Applications*, vol. 16, no. 1, pp. 11–19, oct 2007.

[3] J. Schumann, P. Gupta, and Y. Liu, “Applications of Neural Networks in High Assurance Systems,” in *Neural Networks*, 2010, vol. 268, pp. 1–19.

[4] M. Bozzano and A. Villafiorita, *Design and safety assessment of critical systems*. Auerbach Publications, 2011.

[5] S. Nair, J. L. De La Vara, M. Sabetzadeh, and L. Briand, “An extended systematic literature review on provision of evidence for safety certification,” *Proceedings - IEEE 6th International Conference on Software Testing, Verification and Validation, ICST 2013*, no. February, pp. 94–103, 2013.

[6] A. Kornecki and J. Zalewski, “Certification of software for real-time safety-critical systems: State of the art,” *Innovations in Systems and Software Engineering*, vol. 5, no. 2, pp. 149–161, 2009.

[7] ISO, “26262: Road vehicles-Functional safety,” *International Standard ISO/FDIS*, vol. 26262, 2011.

[8] K. R. Varshney and H. Alemzadeh, “On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products,” *Big data*, vol. 5, no. 3, pp. 246–255, 2017.

[9] K. R. Varshney, “Engineering safety in machine learning,” *2016 Information Theory and Applications Workshop, ITA 2016*, no. 1601.04126v1, 2017.

[10] F. Liu and M. Yang, “Verification and Validation of Artificial Neural Network Models * 3 V & V Approach for ANN Models,” *Springer*, vol. 3809, no. 60434010, pp. 1041–1046, 2005.

[11] P. Koopman and M. Wagner, “Challenges in Autonomous Vehicle Testing and Validation,” *SAE International Journal of Transportation Safety*, vol. 4, no. 1, pp. 2016–01–0128, 2016.

[12] J. Su, D. V. Vargas, and S. Kouichi, “One pixel attack for fooling deep neural networks,” *arXiv:1710.08864*, 2017.

[13] ISO, “ISO/AWI PAS 21448: Road vehicles – Safety of the intended functionality,” 2017.

[14] N. Heess, G. Wayne, D. Silver, T. Lillicrap, Y. Tassa, and T. Erez, “Learning Continuous Control Policies by Stochastic Value Gradients,” *Advances in Neural Information Processing Systems*, pp. 2944–2952, 2015.

[15] K. Golzar, A. Jalali-Arani, and M. Nematollahi, “Statistical investigation on physical-mechanical properties of base and polymer modified bitumen using Artificial Neural Network,” *Construction and Building Materials*, vol. 37, pp. 822–831, 2012.

[16] N. Bahr, *System safety engineering and risk assessment: A practical approach*. CRC press, 1997.

[17] N. Möller and S. O. Hansson, “Principles of engineering safety: Risk and uncertainty reduction,” *Reliability Engineering & System Safety*, vol. 93, pp. 798–805, 2008.

[18] M. Gharib, P. Lollini, A. Ceccarelli, and A. Bondavalli, “Dealing with Functional Safety Requirements for Automotive Systems: A Cyber-Physical-Social Approach,” in *The 12th International Conference on Critical Information Infrastructures Security (CRITIS)*. Springer, 2017.