ARTICLE INFO
Article ID: 12-08-01-0001

doi:10.4271/12-08-01-0001

Letter from the Guest Editors

Focus Issue on Safety of AI-Based Systems

Margriet van Schijndel,[1] Antonio Sciarretta,[1] Olaf Op den Camp,[2] and Bastiaan Krosse[3]

[1]Eindhoven University of Technology, The Netherlands
[2]IFP Energies Nouvelles, France
[3]TNO, The Netherlands

Recent years have shown a rapid increase in technological developments regarding connected and automated (road) vehicles (CAVs). While much progress has been made with step-wise enhancements of merely existing enabling technologies, the complexity of the matter, especially in multi-actor, dynamic traffic situations, has pushed for the sector to advance on artificial intelligence (AI)-based technologies as well. While these have been evolving, generically speaking, much is going on for the specific application in CAVs. For higher levels of automation (SAE Level 3 and beyond), the application of AI technology is indispensable. The full potential and limits of AI in this field are not fully understood yet. There is an eminent need for AI to be explainable, trustworthy, and responsible to enhance user acceptance and safety of such systems when deployed on public roads. In order to trust the decisions made by connected cooperative automated mobility (CCAM) using AI technology, a deeper understanding of the essentials of the control architecture design, based on different performance indicators and beyond the existing testing framework for validation, is needed. Furthermore, methods for assessment and validation of, for example, the operational safety of AI-based systems are in their early stages only. In safety assessment, it is of vital importance to address AI robustness—how the AI responds to a situation that it has not experienced yet [i.e., which is outside its training set, outside its operational design domain (ODD)]—and the specificity of the AI algorithms. For safety assessment of AI-based systems in automated driving, there is a clear need to capture how well the AI behavior is aligned with the required intentions within its ODD, as well as how robust the AI system is in CCAM applications. In advancing to both higher level of automation and higher penetration grades, the safety and robustness of AI for CCAM will play a critical role in user acceptance and uptake, in deployment schedules and type approval, and ultimately, in the successful systemic introduction of these technologies on our roads.

History
Received: 31 Oct 2024
Accepted: 31 Oct 2024
e-Available: 10 Dec 2024

Citation
van Schijndel, M., Sciarretta, A., Op den Camp, O., and Krosse, B., "Letter from the Guest Editors: Focus Issue on Safety of AI-Based Systems," *SAE Int. J. of CAV* 8(1):3–5, 2024, doi:10.4271/12-08-01-0001.

ISSN: 2574-0741
e-ISSN: 2574-075X

For this focus issue, we sought contributions from this field and invited interested authors to submit their research. After a rigorous peer-review process, we selected nine papers for this focus issue.

In the first article, Jörg Bakker et al. discuss the challenges of reliability issues of deep learning algorithms for use in automated driving, as well as in advanced driver assistance systems (ADAS), due to their data-driven and black-box nature. This holds especially true when it comes to accurate and reliable perception of objects in edge case scenarios. So far, the focus has been on normal driving situations, and there is little research on evaluating these systems in a safety-critical context like pre-crash scenarios. The authors describe a project that addresses this problem and provides a publicly available dataset along with key performance indicators (KPIs) for evaluating visual perception systems under pre-crash conditions. The establishment of the dataset is discussed, and an example is detailed of how the evaluation of an AI-based perception system can be performed.

In the second article, Helge Spieker et al. argue that AI methods today are at the center of automated driving and connected mobility, including perception and scene understanding. However, passing control to an AI-based system and trusting its decisions requires the ability to request explanations for these decisions. Societal acceptance of automated driving significantly depends on these AI models' trustworthiness, transparency, and reliability. The authors aim to address this still-open challenge by introducing the qualitative explainable graph (QXG), a unified symbolic and qualitative representation for scene understanding in urban mobility. It enables interpreting an automated vehicle's environment using sensor data and machine learning models. The key advantage of employing a qualitative scene representation lies in introspection and in-depth analysis capabilities.

In the third article, ChungJen Hsu et al. focus on vision-based pedestrian safety applications relying on prompt detection to avoid a crash in various pre-crash scenarios and ODDs. Therefore, verification of training datasets in vision-based vehicle safety applications is crucial to understanding the potential limitations of detection capabilities that may result in a higher safety risk. This article is about the development of a mechanism that detects and tracks pedestrians and cyclists, with subsequently post-processing of the tracking output to identify defined pre-crash scenarios. Tracking of pedestrians and cyclists is performed using open-source object detection and tracking software. The road event awareness dataset is downloaded as the verification target consisting of multiple videos that provide abundant pedestrians (but also cyclists) for training detection and tracking algorithms under various operational design domain conditions.

The following article by Divya Garikapati et al. discusses that due to evolutions in automated vehicle developments, traditional approaches to hard-code control and safety limits into production firmware are no longer appropriate. Especially for machine learning applications, the safety limits need to be adjusted to be flexible and adaptable based on different ODDs and scenarios. The authors extend a previously presented dynamic control limits application strategy using decision-making (DM) engines and a cloud infrastructure. The extended strategy traced the evolution of safety limits, identified scenario classification parameters for the DM engine, and described the primary functions of each autonomy stack layer concerning safety limits application. The classification of safety limits into four categories is presented. Tests on two vehicle platforms are reported, showing the classification of actuation values into the proposed categories using scenario case studies from existing driving logs.

The next article by Jan-Pieter Paardekooper et al. argues that verification and validation (V&V) are the cornerstone of safety in the automotive industry. Automated driving (AD) functionalities pose considerable challenges to the V&V process, especially when based on data-driven AI components. The authors outline and proposed a methodology for V&V of AI-based systems. The backbone of this methodology is bridging the semantic gap between the symbolic level at which the ODD and requirements are typically specified and the subsymbolic, statistical level at which data-driven AI components function.

In the following article, Georg Macher et al. discuss that the automotive domain is characterized by the fact that unexpected and inappropriate responses of in-vehicle systems may lead to safety-critical situations, which might become life threatening. This results in additional requirements for the development, building, and application of AI-driven systems. The article proposes a classification framework for the safety of AI-based systems, in analogy with well-established automotive safety standards. The framework is to be used as a high-level guideline for AI developers in the automotive industry to choose viable pathways to integrate AI-based systems effectively in potentially safety-critical applications while maintaining safety demands.

Further, K.Y. Prashanth et al. present a secure semantic segmentation approach for use in advanced perception systems. Accurate perception is crucial for autonomous driving and is usually obtained with a combination of sensors and machine learning algorithms. The latter are typically used for semantic segmentation, a task aimed at assigning predefined class labels (such as tree, road, etc.) to each pixel of an image. Any security attack on this segmentation system results in falsely segmented pixels with wrong object classes and, thus, directly affects the safety of the autonomous driving. The authors consider a popular encoder–decoder segmentation deep neural network (U-Net), which is vulnerable to these attacks in its last fully connected layer. Hence, a secure semantic segmentation approach is proposed, namely,

the cryptographic security mechanism. The performance of the proposed system is evaluated for the Cityscapes dataset, with the secure semantic segmentation showing satisfactory precision.

The next contribution by Ilma Okanovic et al. dives into the analysis and categorization aspects of AI models and their operation, based on the growing demand for explainable and predictable model behavior. The authors specifically highlight and review the use of explainability in state-of-the-art AI-based scenario understanding and motion prediction methods, which represent an integral part of any AD system. This is broken down into three key axes: the data, the model architecture, and the loss optimization. For each of the axes, the authors outline the general methodologies for introducing explainability and reference and reviewed some practical realizations for each methodology.

In the final article, Paola Natalia Cañas et al. discuss the issue that the application of AI does not only come with enthusiasm, but also with uncertainties, which results in questions and concerns about the functioning of systems that use AI. To enhance the acceptability and explainability of AI systems for CCAM applications, the building and application of AI needs to be reported in a transparent way. The article proposes a set of reporting documents to enhance such transparency. The authors designed a model card tailored to the specific characteristics of AI in CCAM applications, to be used as a guideline for AI developers for reporting. With the model card, the developers are able to fulfill the transparency requirements. The authors indicate that standardization of the reporting processes needs to be established to further enhance transparency toward all stakeholders.

This focus issue on Safety of AI-based Systems includes diverse inputs on new safety developments for several types of AI technologies supporting connected and automated mobility: their trustworthiness, explainability, validation and verification, transparency, and ultimately user and societal acceptability of these essential building blocks. While not aiming to cover all aspects of this growing domain, we hope this selection of articles provides solid insights into both challenges and potential solutions for the safety of AI-based systems for CCAM.

We hope that the reader will be inspired by the different technical research and selected articles.

Finally, we would like to express our sincere appreciation and gratitude to all the authors who made the publication of this focus issue possible, as well as to Colleen Franciscus (Managing Editor at SAE International) for her professionalism and support during the preparation of this focus issue. We furthermore would like to thank Daniel Watzenig, co-Editor-in-Chief, who gave us (and the organization we represent) this opportunity and for encouraging us to detail and edit this focus issue. In addition, our gratitude goes to EARPA, the European Automotive Research Partners Association, for enabling us to have focused discussion and joint learnings on this topic.

Guest Editors
Margriet van Schijndel
Eindhoven University of Technology,
The Netherlands

Antonio Sciarretta
IFP Energies Nouvelles, France

Olaf Op den Camp
TNO, The Netherlands

Bastiaan Krosse
TNO, The Netherlands

Find the online focus issue here.