The Interaction between “Artificial Intelligence” and “the Safety and Security of Cyber Physical Systems”
AI化するクルマにおけるサイバーリスクと安全対策の動向

Paul WOODERSON

Dr. David WARD

NAKANISHI Hideki
中西 秀樹

AI is already being used in vehicle infotainment systems such as car navigation. Recently, one of the most frequently encountered uses of AI in vehicles is in driver support (ADAS) and automated driving systems (ADS). This article explores the use of different forms of artificial intelligence (AI) in cyber-physical systems, in particular automotive systems, the potential safety and security risks posed by AI and how these risks are being addressed through emerging regulations, standards and technical solutions. We examine the functional safety implications of AI-based systems and how this is being addressed by the newly published ISO/PAS 8800. We also provide an overview of typical cybersecurity threat scenarios and attack methods that affect AI-based systems and how those can be mitigated. Finally, we review some of the opportunities that AI can offer for improving functional safety and cybersecurity.

クルマのナビゲーションなどの車載インフォテインメントシステムにおいては，AIがすでに活用されている。昨今，運転支援機能や自動運転機能へもその活用幅が広がっている。本稿では，英国に拠点を持つHORIBA MIRAの取り組みとしてサイバーフィジカルシステム，特に自動車システムにおける様々な形態の人工知能（AI）の活用，AIがもたらす潜在的な安全性とセキュリティへのリスク，そしてこれらのリスクが新たな規制，標準，そして技術ソリューションによってどのように対処されているかについて考察する。AIを用いたシステムの機能安全との関係，これが新たに発行されたISO/PAS 8800でどのように明示されているか解説する。また，AIを用いたシステムに影響を与える典型的なサイバーセキュリティの脅威シナリオと攻撃手法の概要，そしてそれらを軽減する方法についても説明する。最後に，機能安全とサイバーセキュリティの向上に繋がるAI活用の可能性をいくつか検証する。

Introduction and Background

Artificial intelligence (AI) is emerging as a key new technology in modern engineering, with applications in many industry sectors, including in cyber-physical systems such as vehicles.

The benefits of AI come from the unique way it operates. For example, instead of using fixed algorithms based on physics or classical statistics, machine learning uses vast sets of training data to teach an algorithm to identify patterns or specific objects. Once trained, machine learning algorithms can spot these patterns with greater speed and accuracy compared to traditional techniques, whether the task is interpreting human speech for voice control applications or predicting the path of an oncoming vehicle. Compared to traditional software decision-making, AI can accomplish tasks that would otherwise be impracticable with the time or computing power available.

Along with the significant opportunities presented by AI, there are a number of risks, including risks to safety and risks of intentional attacks. The complexity of AI-based systems means that there are many different variables to consider, increasing the potential for failures or vulnerabilities.

Uses of artificial intelligence in vehicles

Applications for AI systems in vehicles range from the use of machine learning for object detection in automated driving systems through to battery life predictions in electric vehicles. Even infotainment systems are starting to adopt generative AI technology.

One of the most frequently encountered uses of AI in vehicles is in driver support (ADAS) and automated driving systems (ADS). AI is seen as attractive since it enables mimicking of human behaviour and human decision making rather than following a rigid set of rules as would be found in a system based on algorithmic decision making. AI is therefore frequently used to interpret sensor data and make decisions based on the perceived traffic environment.

However there are many potential uses of AI beyond ADAS and ADS. One particular example is in battery management systems, where AI can be used to help implement more accurate predictions of battery life both in terms of improving the energy efficiency of individual journeys, but also managing the through-life efficiency and longevity of the energy storage system.

Safety and security risks for AI-based systems

Safety risks

The historical position on the use of AI-based systems in safety contexts has been "to not do". This is due to the non-deterministic behaviour of such systems and the closed box nature of their implementation.

By "non-deterministic" we mean that the output is not fully predictable based on a given set of input conditions; the same set of input data might give different results on different occasions. For example as a result of updated learning data the system might behave differently on a subsequent occasion.

By "closed box" we mean that the implementation of AI based systems (particularly around learned behaviour) is not amenable to detailed analysis. In traditional functional safety methods a software-based system would be designed at successive levels until so-called "software units" are identified which can be subject to standalone and detailed analysis and testing. Examples of this in traditional systems would be a function (or small number of functions) in a language such as "C" that contains a relatively small number of lines of source code.

From a safety perspective the ultimate goal is to demonstrate "assurance" in a system – that it operates correctly in its given operational context and does not present an unreasonable level of risk to people exposed to its behaviour. Traditional methods need adapting for the complexities and uncertainties associated with AI-based systems so that their benefits can be delivered while managing the additional risks associated with them.

Security risks

AI, and in particular machine learning, introduces a number of security-related risks, with various attack methods threatening different parts of the machine learning implementation. Different types of AI threat and examples of attacks that can realise those threats are shown in Figure 1, which extends the classes of threat defined by the Linux®*1 Foundation "Adversarial Robustness Toolbox" project[1].

Due to the learning aspect of Machine Learning systems, attackers could manipulate the behaviour of the system by

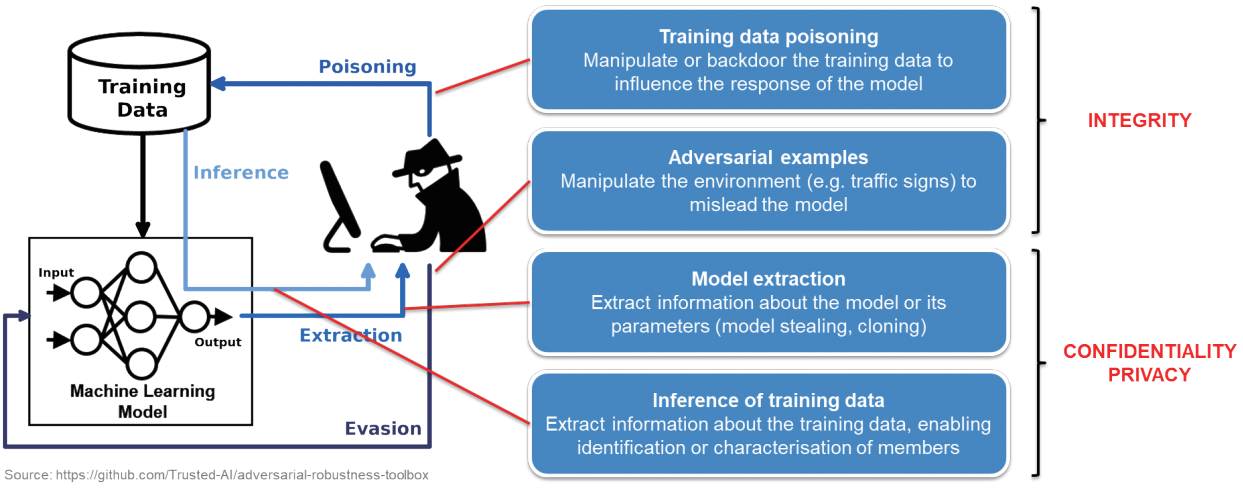


Figure 1 Classes of threat based on the Linux®*1 Foundation "Adversarial Robustness Toolbox".

*1 Registered trademark or trademark of Linus Torvalds in Japan and other countries.

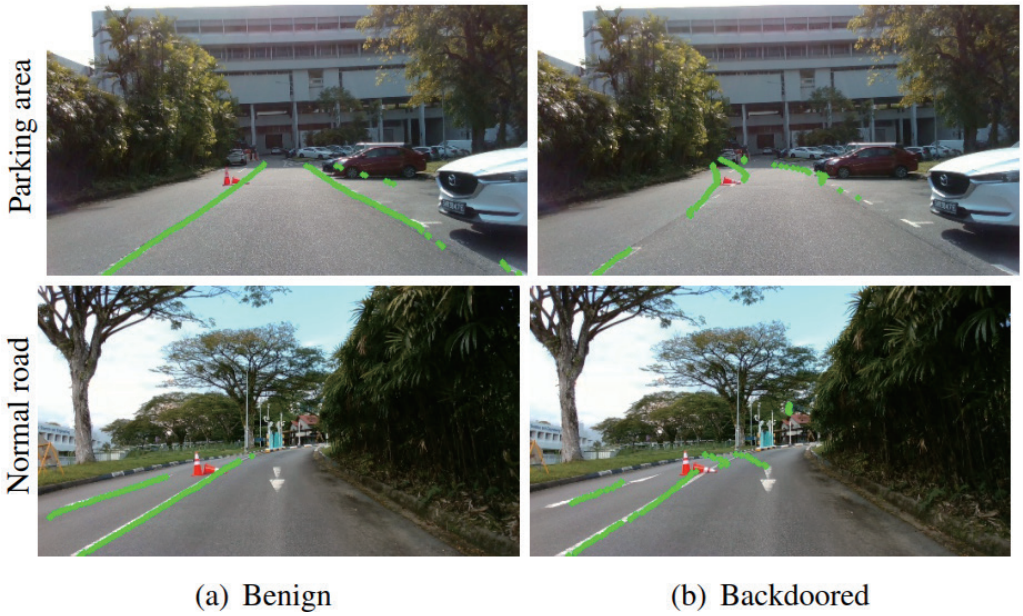


Figure 2 Example of backdoor poisoning attacks (Source: [3])

targeting the training data without needing access to the algorithm itself. These systems are only as reliable as the training data used for the learning process, so removing data or inserting false information into the training data (known as data poisoning) could have a significant impact on the correct behaviour of the algorithm. Research experiments have demonstrated that so-called backdoor poisoning methods could use specific patterns or inputs to bypass normal detection methods[2] . Recent research has demonstrated the practicality of such attacks in the physical world, specifically targeting lane detection systems[3], as illustrated in Figure 2. Experiments have shown these attacks to be effective and robust against various defence solutions, posing significant risks to the security and safety of vehicles and passengers.

With varying degrees driver assistance now seen on many new vehicles, attackers no longer need direct access to a vehicle to influence its operation. Various methods of 'spoofing' to intentionally mislead the vehicle's sensors have been demonstrated by researchers, including the class of attacks known as adversarial examples: an example being small stickers placed on road signs by an attacker (as shown in the image below from research by

Figure 3 Example of adversarial examples attack (Source: [4])

Eykolt et al)[4] which change the way the sign is interpreted by image recognition systems but would be ignored by the human eye(Figure 3).

The value attached to novel and proprietary AI algorithms could make them a particularly lucrative target for intellectual property theft. If a third party can access the AI system – either through a vehicle or by other means, such as an IT system – then they could potentially extract details of the model or its parameters[5] and use the information to either clone or manipulate the behaviour of the target system.

Table 1 Automotive applicability of machine learning attacks.

Application area	AI-based function	Possible attack
Advanced driver assistance systems (ADAS)	Traffic sign recognition	Adversarial examples
Automated driving systems	Object detection and classification	Adversarial examples, Training data poisoning
EV battery management	Predicting battery health and parameters over lifetime	Training data poisoning, Model extraction
Smart cockpit / user experience	Generative AI infotainment functions, voice assistant	Personal information inference

The applicability of some of these types of attack on automotive Machine Learning systems is summarized in Table 1.

A major challenge is that securing any system is an uneven contest between attackers and defenders. Vehicle manufacturers have a finite amount of time and resources available to identify, analyze and resolve security issues before a new vehicle is signed off for production. In contrast, attackers have unbounded opportunity to attack the vehicle throughout its operational lifetime, and they only need to be successful once to cause major damage. It should also be noted that AI techniques may also be harnessed by attackers to increase their advantage and adapt to evolving cybersecurity controls.

Emerging regulations and standards

General AI regulations and standards

General-purpose AI regulatory frameworks are now emerging, such as the EU AI Act and the US National Artificial Intelligence Initiative Act. For specific industry verticals, such as automotive, it is possible that future requirements for approval of AI-based systems in vehicles may be added to existing automotive-specific type approval processes. Alternatively, it may be left to the industry to follow the general-purpose frameworks.

In addition, a major standardization activity is underway in ISO/IEC to develop a range of AI-related international standards. This activity is mainly under the ISO/IEC/JTC 1/SC 42 sub-committee, although other ISO/IEC committees and the European standards bodies CEN/CENELEC are also developing AI standards. These standards cover a wide range of aspects related to AI, from defining governance frameworks, to specific aspects such as ethical considerations. Two examples of these general AI standards that are also important for the safety and security of AI are:

- ISO/IEC 42001 AI Management Systems – establishes a framework for managing the lifecycle of AI systems, ensuring their ethical, reliable, and secure deployment across industries, including automotive. This standard emphasizes governance and leadership commitment, robust risk management, compliance with legal and ethical standards, and stringent data management practices. It outlines best practices for the design, development, deployment, and continuous improvement of AI systems, promoting transparency, accountability, and stakeholder engagement.

- ISO/IEC 22989 AI concepts and terminology – provides definitions of key concepts and terminology related to artificial intelligence. This standard aims to establish a common language for discussing AI across various domains and industries, promoting clarity and consistency in communication. By establishing a clear and consistent vocabulary, ISO/IEC 22989 facilitates better communication, a shared understanding of AI technologies, models, and processes, and collaboration among automotive manufacturers, technology developers, regulators, and other stakeholders.

AI security regulations and standards

Standards specifically regarding the interaction between AI and cybersecurity are currently under development, including a CEN/CENELEC standard on “Cybersecurity specifications of AI systems” and ISO/IEC 27090 “Cybersecurity — Artificial Intelligence — Guidance for addressing security threats to artificial intelligence systems”, which is being developed by ISO/IEC/JTC 1/SC 27. At the current time, these standards are in the early stages of development.

Regarding vehicle specific AI security regulation, with the introduction of UN Regulation 155, cybersecurity is already part of the Vehicle Type Approval process in markets such as the EU, Japan and South Korea. There are currently discussions in the UNECE WP.29, which develops vehicle type approval regulations, about possible future regulation of AI-based systems in vehicles, although at the current time the content and timing of any future regulation is unclear.

AI safety standards

As noted previously, the view on use of AI in safety-related systems is moving from “do not” to “how can we?” There is extensive work taking place to develop a framework of standards for AI. As well as the more general framework described above, the following should be noted with particular reference for safety:

- ISO/IEC TR 5469[6] – this was an initial document prepared in the context of the generic functional safety standard IEC 61508 to provide preliminary information on some of the methods and processes available to incorporate AI elements in safety-related systems,

- ISO/PAS 8800[7] – this builds on ISO/IEC TR 5469 and specifically addresses the use of AI elements within automotive systems. It has a particular focus on the risk associated with undesired safety-related behaviour at a vehicle level due to output insufficiencies, systematic errors and random hardware errors of AI elements. This document also assumes application in the context of ISO 26262[8] – it supplements it rather

than replaces it. Therefore a safety-related system should be developed according to ISO 26262. The functional safety design would create an architecture for the system and allocate safety requirements to its elements – if those elements are implemented using AI then ISO/PAS 8800 would then be applied. There are many important aspects of ISO/PAS 8800 but a significant one is the use of AI properties to help define an assurance claim that use of an AI element achieves absence of unreasonable risk.

- ISO/IEC TS 22440 – currently under preparation, this will build on ISO/IEC TR 5469 and incorporate concepts from ISO/PAS 8800 as industries move towards a "state of the art" on how to use AI within safety-related systems.

Opportunities for AI to improve safety and security

AI provides opportunities to address the asymmetry between attackers and defenders. Technologies such as machine learning are well-suited to identifying anomalous behaviour that may be the first warning signs that a system has been tampered with. The ability to efficiently process huge datasets also makes machine learning a powerful tool for monitoring the diverse and unstructured corpus of published information about new threats, attacks and vulnerabilities. For example, techniques like natural language processing can be used to extract intelligence from unstructured, text-based information, filter out irrelevant content and highlight the potential threats. This provides cybersecurity analysts and engineers with actionable information to support decision making during engineering, vulnerability management and incident response activities.

Conclusion and outlook

In this article we have introduced different applications of AI in automotive cyber-physical systems, including driver assistance systems, automated driving and electric vehicle battery management. We explored the safety risks due to the "non-deterministic" and "closed box" nature of AI-based systems and the need for traditional methods of providing assurance to be adapted to address these additional risks. We have similarly outlined security risks of AI-based systems, including poisoning attacks, adversarial examples and model extraction. Finally, we have introduced some key regulations and standards applicable to AI-based cyber-physical systems, including those covering AI in general, as well as standards specifically addressing the safety and security of AI.

Whatever shape future regulations may take, AI will have a major role to play in the automotive industry and in other industry sectors developing and deploying cyber-physical systems. Organisations and individuals will need to be mindful of the potential risks introduced by AI-based systems, and the new methods required to provide assurance that the risks are appropriately managed. However, with sufficient assurance, AI can open up a whole range of exciting new possibilities to enhance the capabilities of future cyber-physical systems.

HORIBA MIRA is at the forefront of the development of methods safety and security assurance, including initiatives to extend these methods to AI-based systems as described in this article. This expertise helps us provide consulting, test and assurance solutions to automotive and other customers, as well as helping to assure future HORIBA products incorporating AI-based systems.

* Editorial note: This content is based on HORIBA's investigation at the year of issue unless otherwise stated.

References

[1] Linux Foundation, 'Adversarial Robustness Toolbox', [online] Available at: https://github.com/Trusted-AI/adversarial-robustness-toolbox, last accessed 2 May 2025.

[2] Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C. and Li, B. 2021, 'Manipulating machine learning: Poisoning attacks and countermeasures for regression learning', arXiv preprint, arXiv:1804.00308

[3] Han, X., Xu, G., Zhou, Y., Yang, X., Li, J. and Zhang, T. 2022, 'Physical backdoor attacks to lane detection systems in autonomous driving', Proceedings of the 30th ACM International Conference on Multimedia (MM '22), Association for Computing Machinery, New York, NY, USA, p. 2957–2968.

[4] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning visual classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 1625–1634.

[5] Nightfall AI, 2024. Training Data Extraction Attacks: What is a Training Data Extraction Attack? [online] Available at: https://www.nightfall.ai/ai-security-101/training-data-extraction-attacks#what-is-a-training-data-extraction-attack, last accessed 2 May 2025.

[6] ISO/IEC TR 5469:2024 Artificial intelligence — Functional safety and AI systems, Edition 1, January 2024.

[7] ISO/PAS 8800:2024 Road vehicles — Safety and artificial intelligence, Edition 1, December 2024.

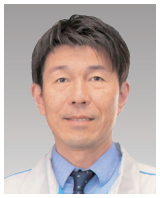
[8] ISO 26262:2018 Road vehicles — Functional safety, Edition 2, December 2018.

Paul WOODERSON

Chief Engineer – Cybersecurity
Vehicle Resilience,
Cybersecurity and EMR
HORIBA MIRA

Dr. David WARD

Global Head of Functional Safety
Functional Safety
HORIBA MIRA



NAKANISHI Hideki

中西 秀樹

Consultancy & Applied Solutions(CAS) Lead
Group Strategy Division,
Energy & Environment Task Force
HORIBA, Ltd.