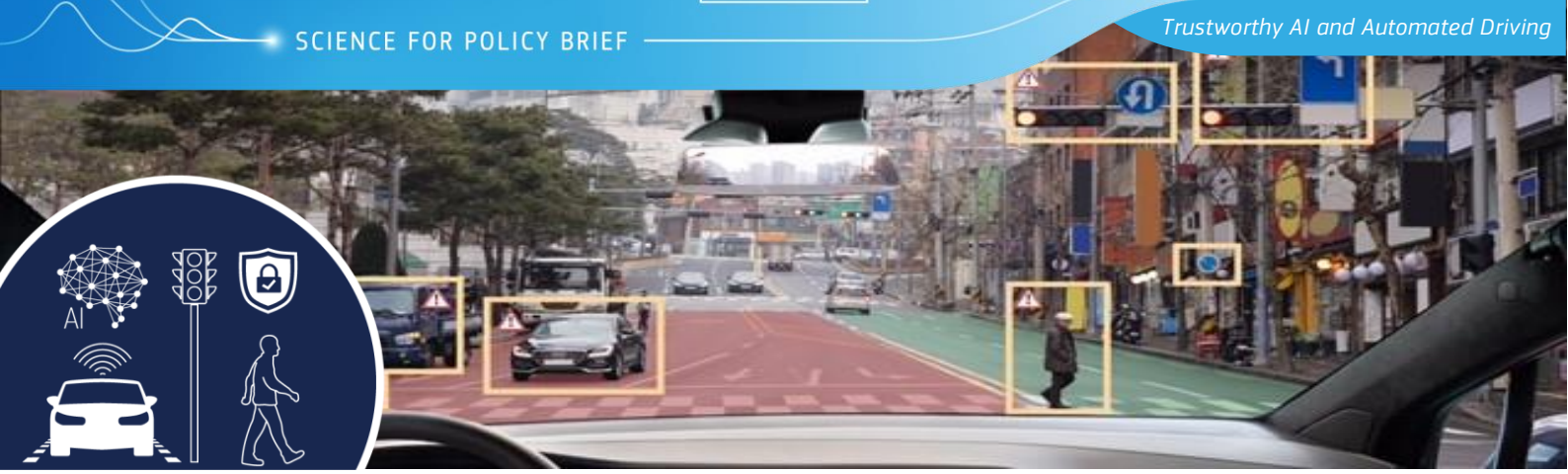
European Commission

SCIENCE FOR POLICY BRIEF

Trustworthy AI and Automated Driving



Artificial Intelligence in Automated Driving: an analysis of safety and cybersecurity challenges

HIGHLIGHTS

- → Artificial Intelligence (AI) is a powerful technology that will play a central role in mobility of the future.
- → The breakthroughs achieved by AI systems in automated driving are contrasted by their higher complexity and opacity, potentially leading to safety and cybersecurity risks.
- → The growing digitalisation of vehicle systems increases the potential cybersecurity attack surface and can lead to stronger impacts if automation mechanisms are compromised.
- → Biases in and lack of generalisation of AI systems and data may cause vehicles to malfunction in natural conditions or as a result of a cyber-attack, putting the lives of passengers and road users at risk.
- → Vehicle testing procedures need to take into account the specificities of AI to ensure that safety and cybersecurity risks are properly addressed.
- → Addressing by design the AI safety and cybersecurity challenges is key to securing the many benefits that automated driving can bring to society.

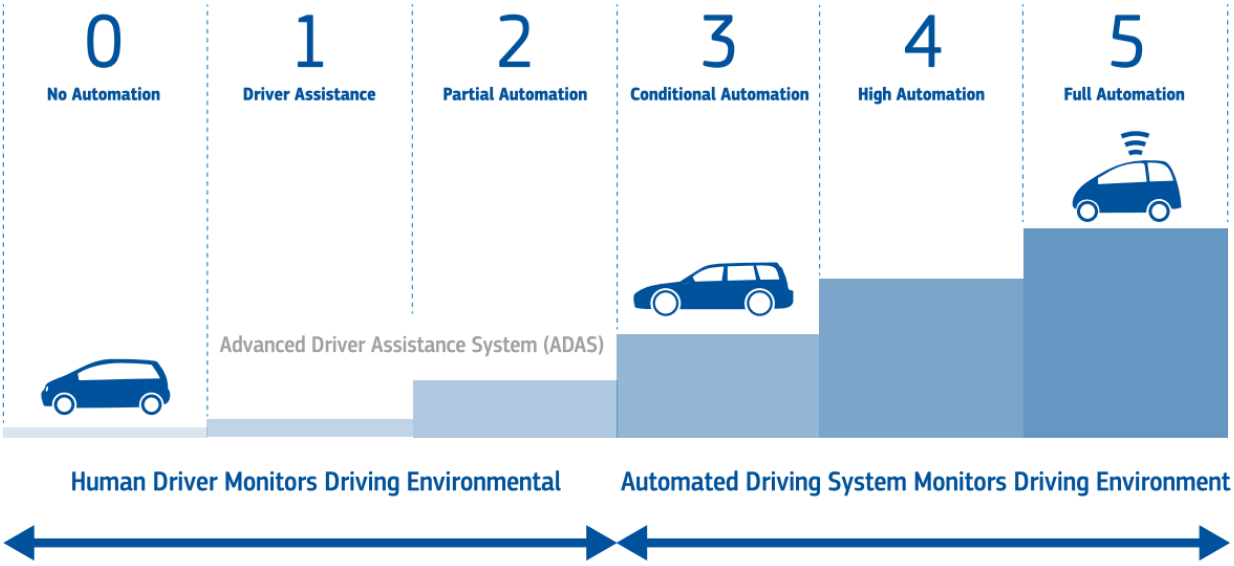
INTRODUCTION

Modern vehicles are increasingly equipped with automation mechanisms that are designed to assist or replace human drivers and enhance the safety of road users. This trend leads to the development of a new generation of automated vehicles, such as cars, trucks, or buses, that are capable of circulating in the public with limited or no human intervention.

While such capabilities were out of reach only a decade ago, recent advances in AI have triggered a positive outlook for this technology. Level 4 and 5 automated vehicles (see Figure 1) are being actively developed by technological companies and car manufacturers, despite uncertainty about future commercialisation.

Joint Research Centre

Figure 1 – The SAE J3016 standard defines six levels of driving automation for on-road vehicles. Regarding automated vehicles, Level 4 describes vehicles able to autonomously (i.e. without any human driver intervention) perform all driving functions under certain conditions (e.g. on a given type of roads), whereas Level 5 describes vehicles able to autonomously perform all driving functions under all conditions.



Source: JRC

Vehicles of lower levels are already available on the market, providing Advanced Driver-Assistance Systems (ADAS), some of them based on Artificial Intelligence (AI). Even if these assistance functions are expected to be operated under human supervision, they should still provide sufficient guarantees in terms of safety and cybersecurity.

This brief

- → reflects on the evolution of safety and security testing of automated vehicles, taking into account the limitations of AI technology, in particular machine learning based systems,
- → provides an outlook of the challenges for road safety and security introduced by the adoption of automated features powered by AI in vehicles,
- → explains the novel threats introduced by AI technology, based on recent scientific developments,
- → justifies the need for an evolution of testing methodologies to properly assess such threats.

To assess the threat landscape, the discussion is divided according to the nature of the risks, making a distinction between situations where malfunctions appear in natural driving conditions, and situations where they are deliberately caused by an adversary.

By its nature, automated driving constitutes a high-risk application: malfunctions and cyberattacks can cause safety problems and harm in the physical world, potentially at large scale. Automated vehicles are cyber-physical systems[1] operating in particularly challenging environments, designed by and for humans. They are equipped with advanced perception and planning systems capable of recognising road markings (lane, signs, etc.), users (vehicles, cyclists, pedestrians, etc.) and objects, and of taking actions (accelerating, braking, turning, etc.,) in order to achieve the desired trajectory. These capabilities largely rely on complex digital systems and software powered by AI technologies, where machine learning[2], and in particular deep learning[3], plays a core role.

These techniques, combining mathematical modelling, advanced algorithms, large volumes of data, and tremendous computational power, are steadily closing the gap between automated systems and human reasoning capabilities in the narrow context of automated driving. Machine learning techniques are becoming a key enabler in perception and planning tasks, sensing and making sense of the environment, and determining the trajectory of the vehicle and the actions (e.g. accelerating, turning, etc.) to achieve this trajectory, while ensuring safe and secure driving aligned with human values.

Even though machine learning has reached unprecedented levels of performance in recent years, it has limitations in terms of understandability and robustness that can make systems subject to various vulnerabilities and issues[4]. It is important to understand the threats introduced by AI technologies, and to integrate their specificities in testing procedures in order to ensure the compliance of AI systems in

[1] https://ec.europa.eu/digital-single-market/en/cyber-physical-systems
[2] Bishop, 'Pattern Recognition and Machine Learning', 2006, Springer.
[3] Goodfellow et al., 'Deep Learning', MIT Press, 2016.
[4] Rudner and Toner, 'Key Concepts in AI Safety: Robustness and Adversarial Examples', 2021.

automated vehicles with current safety and security requirements[5].

Addressing these issues will be crucial for the adoption of automated vehicles by consumers, in order to create trust in such novel, highly technological products.

Policy context

The challenges raised by automated driving led to the development of several regulations and initiatives aiming to ensure safe and secure vehicles. The European Commission launched several initiatives, such as the Cooperative Intelligent Transport Systems (C-ITS) deployment platform[6], with the objective to identify and agree on how to ensure interoperability of intelligent transport across borders and along the whole value chain, and the C-Roads platform[7] to develop and share technical specifications to verify interoperability through cross-site testing.

The Commission also published a Strategy on Cooperative Intelligent Transport Systems to facilitate the convergence of investments and regulatory frameworks across the EU, and a Strategy for mobility of the future setting out specific actions to implement a pilot on common EU-wide cybersecurity infrastructures and processes, needed for secure and trustful communication between vehicles and infrastructure. In 2019, the European Commission set up a Commission Expert group on cooperative, connected, automated mobility (CCAM) to provide advice and support to the Commission in the field of testing and pre-deployment activities.

In parallel, specific actions regarding the uptake of AI in automated driving have been taken, such as the publication in 2020 of a report on the Ethics of Connected and Automated Vehicles[8] by an independent group of experts that includes 20 recommendations covering dilemma situations, the creation of a culture of responsibility, and the promotion of data, algorithm and AI literacy through public participation. In February 2021, a joint ENISA-JRC report[9] on the cybersecurity risks of AI in automated driving was published, describing the potential vulnerabilities of AI components embedded in vehicles, and providing recommendations to mitigate them.

From a general standpoint, ensuring safety and security of AI systems is a cornerstone of recent policy initiatives of the European Commission, as part of a commitment to establishing a set of principles for trustworthy AI. This led to the publication of a proposal for a regulation of AI in April 2021[10], putting forward legislation addressing the human and ethical implications related to the uptake of the technology in end-user products and services. In particular, it details specific high risk cases, defined by sectors (transport, healthcare, etc.) and the impact (e.g. physical harm, damage, etc.), for which a set of requirements, including robustness and cybersecurity, has to be met by the system providers.

Automated vehicle functions are regulated by the United Nations Economic Commission for Europe (UNECE) as specific scenarios such as advanced emergency braking (AEB)[11], lane keeping assist (ALKS)[12], and lane departure warning (LDW)[13]. Cybersecurity of vehicles in general has also been regulated under UNECE[14], with various measures to minimise risks throughout the entire life cycle of a vehicle. Current testing is based on black-box approaches to ensure minimum levels of safety, but does not cover all principles for trustworthy AI, which may lead to unexpected accidents (see Box 1). In this

Box 1: Malfunction of automated vehicles

Accidents involving AI features of level-2 automated vehicles have occurred multiple times all over the world. In 2018 in the United States, the autopilot feature of a vehicle was not able to detect a white truck crossing the road in twilight conditions, leading to a crash. In 2020 in Taiwan, an automated vehicle crashed into an overturned truck on the highway. Preliminary studies suggest that the cameras may have mistaken the white roof of the truck without wheels for an overexposed portion of the scene, while the low resolution of radar may not have been able to locate the object on the same lane as the vehicle. It has been noted that the drivers may have relied excessively on automated features, and were not in a position to recover manual control of the vehicle in time.

Fully-automated test vehicles have also been subject to accidents. In 2018, a vehicle struck a pedestrian while driving autonomously. The investigation concluded that the system only succeeded in detecting the pedestrian and predict her path a few seconds before impact, therefore preventing the operator from using the emergency brake.

Sources

- Collision Between Car Operating with Partial Driving Automation and Truck-Tractor Semitrailer, Delray Beach, Florida, 2019', National Transportation Safety Board,
- Accident report NTSB/HAB-20/01, 2019.
- *https://www.forbes.com/sites/bradtempleton/2020/06/02/tesla-in-taiwan-crashes-directly-into-overturned-truck-ignores-pedestrian-with-autopilot-on/*
- Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018', National Transportation Safety Board, Accident report NTSB/HAR-19/03, 2019.

[5] Berghoff et al., 'Towards Auditable AI Systems', 2021.
[6] *https://ec.europa.eu/transport/themes/its/c-its_en*
[7] *https://www.c-roads.eu/platform.html*
[8] Horizon 2020 Commission Expert Group to advise on specific ethical issues raised by driverless mobility (E03659), 'Ethics of Connected and Automated Vehicles: Recommendations on Road Safety, Privacy, Fairness, Explainability and Responsibility', 2020.
[9] ENISA-JRC, 'Cybersecurity Challenges in the Uptake of Artificial Intelligence in Autonomous Driving.', 2021.
[10] European Commission, 'Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence', 2021.
[11] UN ECE R131, EUR-LEX 2014 L 214/47
[12] UN ECE R 157 EUR LEX 2021/389
[13] UN ECE R130 EUR LEX 2013 L 178/29
[14] UN ECE R155 EUR LEX 2021/387

respect, AI-based functions involved in automated driving will likely require an adaptation of current regulations in line with a future regulation on AI at the European level, to ensure proper assessment before deployment, in particular regarding their accuracy, robustness, and cybersecurity.

AI SAFETY FOR AUTOMATED DRIVING

AI safety in automated vehicles aims to identify potential causes of failures of AI systems, and reduce the likelihood of the occurrence of unintended behaviours[15]. To ensure the protection of human life and the security of public spaces, AI safety considerations should be properly addressed in the development and deployment of automated capabilities in vehicles and, more broadly, in software used in vehicles for automated driving. This requires taking into account the reliability of each individual subcomponent focused on a dedicated task (e.g. traffic sign detection, trajectory prediction and planning, etc.), as well as of the system as a whole.

Three key areas of AI safety are:

- specification - ensuring that the behaviour of systems is fully aligned with the intentions of the designers,
- robustness - providing guarantees that systems do not display unexpected behaviour in the wide range of conditions they may encounter,
- assurance - ensuring that systems are auditable and understandable by human supervisors.

Safety considerations vary according to the nature of the task the AI system aims to solve, and the techniques that are employed. The robustness aspect is particularly crucial in the context of automated driving. The outer environment in which automated vehicles evolve is made of an uncountable number of situations that are only partially captured by the data sets on which systems are developed. Variability factors include, amongst others, weather, lighting, road infrastructures, types of vehicles, vegetation, buildings, road behaviours, roadway conditions, etc. The impact of failures in driving may be very serious (e.g. life loss, injuries, damages, etc.) considering the driving settings. In particular, a vehicle's high speed may not allow for a human driver to intervene and take back control in time when the AI based decision-making systems cannot reliably perform a task (see Box 1).

The evaluation of the robustness of AI systems should also take the geographical context into account in which the development of automated vehicles takes place: a significant share of the research and innovation activities on AI in automated driving takes place in the US and in China, raising concerns about the reliability of systems trained on datasets that may not reflect the specificities of EU roads (e.g. multilingual signs, narrow roads, specific driving rules and behaviours, etc.).

Finally, inspecting the inner mechanisms at play in AI-based decision-making processes is crucial for safety considerations. It is important to ensure, before deployment, that AI systems display a sufficient level of safety to drive on open roads, that systems are properly audited, and that liability can be determined in case of an accident. However, the opaqueness of AI systems, even for experts, is a major hindrance to the understanding of the logics involved in the decision taken by automated systems, limiting the capacity to understand the decisions taken by vehicles.

AI CYBERSECURITY FOR AUTOMATED DRIVING

Modern vehicles come equipped with an increasing number of digital systems and online connectivity to offer smart functionalities and seamless integration within the wide array of digital services used by citizens. This growing digitalisation of vehicles results in a larger attack surface[16] that can be exploited by malicious actors driven by profit (e.g. using ransomware) or by intent to cause physical harm. Automated vehicles need therefore to be properly secured to mitigate the safety risks caused by malicious actions, due to their integration in a high-risk environment.

The increasing uptake of AI technologies in vehicles amplifies this trend, bringing an additional layer of complexity on top of classical software components[17]. This more complex and larger digital ecosystem in vehicles raises significant cybersecurity challenges, particularly considering the potential impact a cyberattack could have if the core functionalities of the vehicle are compromised. In 2014, security researchers demonstrated for the first time how the digital systems of a well-known vehicle model sold worldwide could be attacked over the Internet to take remote control over the vehicle to cause crashes[18].

The inclusion of more advanced AI components in higher levels of automation further complicates this picture by introducing a whole new range of potential vulnerabilities, the malevolent exploitation of which could cause intended offence and harm. Recent years have seen an increasing amount of research and practical examples highlighting new attacks against machine learning systems[19] [20] [21].

[15] Arnold and Toner, 'AI Accidents: An Emerging Threat', 2021.
[16] The term "attack surface" is used to describe the sum of all potential entry points that can be used by an adversary to compromise the security of a system.
[17] Musser and Garriott, 'Machine Learning and Cybersecurity: Hype and Reality', 2021.
[18] Miller and Valasek, 'Remote exploitation of an unaltered passenger vehicle'. Black Hat USA 2015.
[19] Huang et al, 'Adversarial Machine Learning', ACM workshop on Security and artificial intelligence, 2011.
[20] Szegedy et al., 'Intriguing properties of neural networks' International Conference on Learning Representations, 2014.
[21] Carlini and Wagner, 'Towards evaluating the robustness of neural networks', IEEE Symposium on Security and Privacy, 2017.

Box 2: Cybersecurity attacks against automated vehicles

Adversarial attacks happen when an adversary alters the inputs of the AI system to deceive its decision-making process. This can be done digitally, e.g. by adding a perturbation on the data stored in memory, or by altering the environment captured by the sensors of the vehicle (see Figure 2).

In 2019, the Chinese cybersecurity expert group Keen Security Lab of Tencent published a widely recognised experimental security research of a commercial driving assistance system present in vehicles with various level-2 autonomous driving functionalities. This system, based on AI technologies, is embedded in the ICT infrastructure of the car and directly connected to the engine control, while mostly relying on camera sensors.

First, they demonstrate that a malicious actor can remotely enter the ICT system of the car by exploiting vulnerabilities in the system software. After that, the malicious actor is able to perform a number of actions on the classical software, including steering the wheel remotely. In a second stage, they show that it is also possible to attack the AI-based autopilot system, some functions of which being solely based on machine learning perception systems, which make them vulnerable to adversarial attacks. In a range of examples, Keen security researchers produce adversarial examples capable of making the lane detection assistance system turn the car into the reverse lane, only requiring the access to the outputs of deep learning models

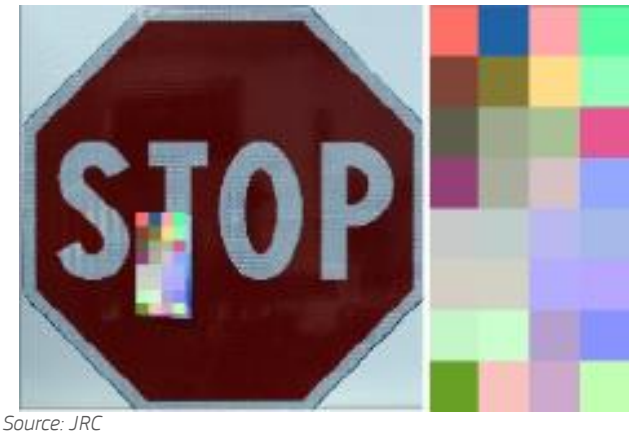
Source
https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf

The main types of attacks include:

- *evasion attacks and adversarial examples* that consist of supplying specifically crafted inputs – adversarial examples - leading to misclassifications or other faulty behaviours,
- *data poisoning* that consists of tampering with the training data of machine learning systems to add vulnerabilities in the fitted AI model (e.g. adding backdoors or inducing a specific malfunction).

Other attacks against the AI model or training data have been designed to extract model parameters or data used in the development phase of models, with potential risks for privacy, trade secrets, and system integrity.

Figure 2 – Adversarial attack against a traffic sign carried out in JRC: a sticker (right) is placed onto a 'Stop' sign (left) causing a traffic sign recognition system to incorrectly classify it as an 'End-of-speed-limit' sign.



Source: JRC

So far, adversarial examples are considered as the main threat for automated vehicles, as most subsystems for perception rely on deep learning based models known to be highly susceptible to this kind of attack. Instances of these attacks may happen in physical context, by altering the environment perceived by sensors. Although such cyberattacks may be hard to implement, examples of successful attempts on commercial vehicles in real world settings have already been demonstrated in controlled environments. This includes for instance deceiving traffic sign recognition systems by placing stickers on signs, leading automated vehicles to accelerate past the speed limit (see Box 2). These kinds of attack can be set up without knowledge of the inner design of AI systems, simply by gaining access to the ICT system and monitoring the outputs of the AI components.

Another prominent threat for automated vehicles comes from data poisoning, as it is very common for AI systems to be updated with new data to make them more accurate over time. Attacking the communication channels between vehicles and manufacturers could allow an adversary to inject corrupted data in the training stage of vehicles, leading to an alteration of the AI components that could be deployed on a large scale through Over-the-Air updates.
Both attacks exemplify the need to consider the full supply chain around AI components when considering the cybersecurity of automated vehicles, and not just AI-specific vulnerabilities (e.g. by securing the way automated vehicles update AI and non-AI software[22]).

[22] UN ECE R156 EUR LEX l 82/60

CONCLUDING REMARKS

Replicating the complex and multifaceted skills of human drivers in automated vehicles is a huge technical challenge, which is progressively being solved by the tremendous advances of AI technologies. Machines rely on statistical patterns in data to understand their environment and infer a decision. As the logical mechanisms are not explicitly defined, a higher degree of complexity and opacity is introduced that can lead to unexpected behaviour. AI systems in vehicles may be susceptible to malfunctions or cyber-attacks that could endanger the life of passengers and other road users.

This brief provides a qualitative analysis of the challenges that follow the uptake of AI components in vehicles in terms of safety and security. It is important to ensure that safety measures, security testing and auditing of software used in automated vehicles account for AI-specific potential vulnerabilities. This will only be possible with the adoption of practices that take into account the nature of AI systems and their supply chain. While AI systems will likely not achieve a perfect correctness, the understanding and control of the uncertainty of AI systems will be crucial to understand how to define sound regulations.

The development of AI systems takes place in a complex supply chain, involving many stakeholders, assets, and complex processes, including data management and feedback loops, that need to be properly tested, secured and integrated in traditional ICT development lifecycles. The establishment of testing strategies accounting for the inherent flaws of AI systems will need to be pursued, in line with the current principles set out in the European Commission proposal for a regulation of AI. The concrete implementation of these principles will require a joint research and development effort in transport, AI, and cybersecurity.

Beyond compliance with safety and security requirements, these efforts are also necessary to ensure the consistency of liability regimes with the emergence of new digital technologies such as AI. Testing will never completely exclude the possibility of accidents or malfunctions involving automated vehicles. The understanding of the sequence of events that led to an accident may require in-depth analysis of the inner mechanisms of the AI components of the vehicle as part of the investigation to determine liabilities. In this context, current liability regimes may need to be revisited[23].

In addition to testing, limitations of AI systems in comparison to human abilities can be addressed by shifting the design of road infrastructures, from an approach exclusively thought out for human drivers to a new generation of infrastructure including signals that can be processed efficiently and reliably by machines. This approach is in line with the C-ITS strategy currently under deployment at European level, which aims to facilitate the convergence of investments and regulatory frameworks across the European Union to connect vehicles and infrastructures. Getting information from the infrastructure in a reliable and secure way may provide redundancy channels that could complement AI based perception and planning systems of the vehicle in particularly challenging environment such as dense urban areas.

AI technologies are key for the development of a new generation of vehicle safety systems that can better protect drivers and pedestrians handling situations where human driver capabilities fail (e.g. for lack of concentration, or during fast-paced events). Vehicles with high levels of automation have also shown their potential to avoid accidents in complex situations that might be challenging for human drivers. However, AI failures, both accidental or as a result of a cyberattack, have also occurred in situations where human drivers would not have been confused.

AI is a powerful technology that will play a central role in mobility of the future. Fully addressing the specific safety and cybersecurity challenges linked to the integration of AI technologies in vehicles is key to securing the many benefits that automated driving can bring to society, ensuring that its uptake results in higher safety levels on European roads.

AUTHORSHIP

This policy brief was prepared by Ronan Hamon, Henrik Junklewitz and Ignacio Sanchez, with contributions from David Fernandez Llorca, Emilia Gomez, Antonio Herrera and Akos Kriston.

DISCLAIMER

The JRC is carrying out research on the challenges of addressing the different requirements of trustworthy artificial intelligence in the context of automated driving. This brief is one of a series of Science for Policy briefs summarising the main outcomes on this research. To date, another brief has been published in this series:

Fernández Llorca, D., Gómez, E. "Artificial Intelligence in Autonomous Vehicles: towards trustworthy systems", European Commission, 2022, JRC128170.

COPYRIGHT

[23] Twigg-Flesner, 'Guiding Principles for Updating the Product Liability Directive for the Digital Age', 2021, https://papers.ssrn.com/abstract=3770796.