enisa
EUROPEAN UNION AGENCY FOR CYBERSECURITY

European Commission

CYBERSECURITY CHALLENGES IN THE UPTAKE OF ARTIFICIAL INTELLIGENCE IN AUTONOMOUS DRIVING

EUR 30568 EN

ABOUT ENISA

The European Union Agency for Cybersecurity, ENISA, is the Union's agency dedicated to achieving a high common level of cybersecurity across Europe. Established in 2004 and strengthened by the EU Cybersecurity Act, the European Union Agency for Cybersecurity contributes to EU cyber policy, enhances the trustworthiness of ICT products, services and processes with cybersecurity certification schemes, cooperates with Member States and EU bodies, and helps Europe prepare for the cyber challenges of tomorrow. Through knowledge sharing, capacity building and awareness raising, the Agency works together with its key stakeholders to strengthen trust in the connected economy, to boost resilience of the Union's infrastructure, and, ultimately, to keep Europe's society and citizens digitally secure. For more information, visit www.enisa.europa.eu .

ABOUT JRC

The Joint Research Centre is the European Commission's science and knowledge service. The JRC is a Directorate-General of the European Commission under the responsibility of Mariya Gabriel, Commissioner for Innovation, Research, Culture, Education and Youth. Our researchers provide EU and national authorities with solid facts and independent support to help tackle the big challenges facing our societies today. Our headquarters are in Brussels and we have research sites in five Member States: Geel (Belgium), Ispra (Italy), Karlsruhe (Germany), Petten (the Netherlands) and Seville (Spain). Our work is largely funded by the EU's budget for Research and Innovation. We create, manage and make sense of knowledge, delivering the best scientific evidence and innovative tools for the policies that matter to citizens, businesses and governments.

For more information, visit https://ec.europa.eu/jrc.

CONTACT

For contacting the authors, please use resilience@enisa.europa.eu or https://ec.europa.eu/jrc/en/contact/form
For media enquiries about this paper, please use press@enisa.europa.eu or jrc-press@ec.europa.eu

AUTHORS

ENISA
Georgia Dede
Rossen Naydenov
Apostolos Malatras

JRC
Ronan Hamon
Henrik Junklewitz
Ignacio Sanchez

ACKNOWLEDGEMENTS

We would like to acknowledge the following experts who have reviewed the report (in alphabetical order):

Christian Berghoff (BSI)
Ernesto Damiani (Universita degli Studi di Milano)
David Fernandez Llorca (JRC)
Tijink Jasja (Kapsch)
Victor Marginean (Continental Automotive GmbH)
Gerhard Menzel (JRC)
Isabel Praça (Instituto Superior de Engenharia do Porto)
Alexandru Vasinca (Cognizant Softvision)

LEGAL NOTICE

COPYRIGHT

JRC122440
EUR 30568 EN
PDF ISBN 978-92-76-28646-2 ISSN 1831-9424 doi:10.2760/551271

Luxembourg, Publications Office of the European Union, 2021

How to cite this report: Dede, G., Hamon, R., Junklewitz, H., Naydenov, R., Malatras, A. and Sanchez, I., Cybersecurity challenges in the uptake of artificial intelligence in autonomous driving, EUR 30568 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-28646-2, doi:10.2760/551271, JRC122440.

TABLE OF CONTENTS

List of figures

EXECUTIVE SUMMARY

New generations of cars are making use of advances in the field of Artificial Intelligence (AI) to provide semi-autonomous and autonomous driving capabilities, achieving a technological breakthrough that will strongly impact existing behaviours and practices. Beyond the undeniable benefits of autonomous driving for many aspects of our societies, the question of the safety and security of this technology, which by definition is intended to operate with limited human supervision, has emerged. The answers provided by regulatory bodies on these issues are likely to play an important role for the adoption of autonomous vehicles (AVs) in society. This is all the more important given that machine learning (ML) techniques, at the core of the AI components developed to mimic human cognitive capabilities, have been proven to be highly vulnerable to a wide range of attacks that could compromise the proper functioning of autonomous vehicles, and pose serious threats to the safety of persons, both inside and outside of a vehicle. In this context, understanding the AI techniques used for autonomous driving and their vulnerabilities in the cybersecurity threat landscape is essential to alleviate the risks and ensure that benefits will not be counterbalanced by stronger safety risks.

Cybersecurity of AVs is classically approached through the angle of the security of digital systems. This is all the more relevant as modern vehicles are fully controlled by electronic components, vulnerable to physical and remote attacks exploiting classical cybersecurity vulnerabilities. With this report however, the objective is to raise awareness about the potential risks connected to the AI components in charge of replicating tasks previously addressed by human drivers, such as making sense of the environment or taking decisions on the behaviours of the vehicle. By their nature, those AI components do not obey the same rules as traditional software: ML techniques are indeed relying on implicit rules that are grounded on the statistical analysis of large collections of data. While this enables automation to reach unprecedented cognitive capabilities, it opens at the same time new opportunities for malicious actors, who can exploit the high complexity of AI systems to their own advantage. Securing such systems requires to consider these AI specific issues on top of the traditional cybersecurity risks connected to digital systems, in the context of the full supply chain involved in their development and of their integration with other automotive systems.

This report aims to provide insights on the cybersecurity challenges specifically connected to the uptake of AI techniques in autonomous vehicles. It starts by describing the dynamic policy context with which this initiative is aligned, at both the European and international levels. Institutional and private actors have been very active to outline the high-level principles and standards that should govern the development of AV, either explicitly, with dedicated automotive guidelines, or through the definition of sets of practices driving the expansion of AI and cybersecurity. In this respect, the European institutions have conducted various initiatives for developing trustworthy AI, where cybersecurity and intelligent transportation play a significant role.

Subsequently, this report delves into the technical aspects of AI in the automotive sector, with the aims to better comprehend the technological concerns of AI, as well as to get a sense of the level of integration of AI in AV. This includes an extended description of the areas in which AI plays a role, to ensure the proper implementation of cognitive capabilities inside automotive systems. Autonomous driving requires addressing a host of smaller subtasks (recognizing traffic signs or roads, detecting vehicles, estimating their speed, planning the path of the vehicle, etc.), each of them trivially performed by humans, but requiring carefully engineered AI systems to automatically address them. AI software components in an AV do not form a monolithic system, but rather rely on a complex combination of large and varied collections of data, themselves obtained by several types of sensors, and a rich set of AI methodologies, based on scientific works from statistics, mathematics, computing, and robotics. Starting from the high-level functions, an extended description of the landscape combining AI techniques, sensors, data types, and cognitive tasks highlights the sheer abundance of approaches and ideas that have made AV a reality. We claim that the understanding of these technical elements in the automotive context is essential to put into perspective their cybersecurity implications of these AI-based components. A mapping of automotive functions to AI techniques is provided to highlight the connections between automotive and scientific concepts, making direct links between automotive functionalities, intermediate subtasks, and ML techniques.

After this technical presentation, a state-of-the-art literature survey on security of AI in the automotive context discusses the main concepts behind the cybersecurity of AI for autonomous cars. Security of AI in general lies outside the scope of this report, and the interested reader is referred to the recently published ENISA AI Threat Landscape [1] to get the full picture on this matter. Instead, a focus is specifically made on adversarial machine learning that regroups a set of techniques that are currently the main approaches susceptible to compromise AI components of AVs. They allow a malicious actor to design specific attacks that could deceive AI systems while staying undetectable by human supervisors. Concretely, carefully crafted patterns can be disseminated in the environment to alter the decision-making process and induce unexpected behaviour of the vehicle. Typical examples include adding paint on the road to misguide the navigation, or stickers on a stop sign to prevent its recognition. Despite the complexity to undertake these kinds of attacks, and in particular to make them undetectable by human eyes, the dire consequences in terms of safety should encourage car manufacturers to implement defence mechanisms to mitigate these type of AI risks. The description of these attacks, which may not necessarily require access to the internal system of the vehicle, is accompanied by real-world cases involving autonomous or semi-autonomous cars fooled by attackers. This is subsequently illustrated, both theoretically and experimentally, by realistic attack scenarios against the AI components of vehicles, extending the discussion to other types of vulnerabilities of AI.

In conclusion of this report, a set of challenges and recommendations is provided to improve AI security in AVs and mitigate potential threats and risks. This is motivated by the importance of relying on the pillars that have been at the core of cybersecurity methodologies developed along the years for traditional software, while at the same time taking into account the particularities of AI systems. In light of the connections between AI and AVs brought forward in this report and their consequences in terms of security, the following recommendations are put forward.

Systematic security validation of AI models and data

Data and AI models play an important role in the implementation of autonomous capabilities in AVs. These components are dynamic in nature and can change their behaviour overtime as they learn from new data, are updated by manufacturers, or encounter unexpected or intentionally manipulated data. This requires that the security and robustness assessments of AI components do not just take place at a given point in time during their development, but instead are systematically performed throughout their lifecycle.

This systematic validation of both AI models and data is essential to ensure the right behaviour of the vehicle when faced with either unexpected situations or malicious actions such as attacks based on the alteration of inputs, including poisoning and evasion attacks. This implies developing and maintaining strict continuous processes to make sure that data that are used at the development and production stages have not been altered with a malicious intent, and that models do not contain vulnerabilities that could be exploited. It also means that links between industrial actors and research centres have to be reinforced to address the challenges associated with the implementation of this systematic validation.

AI supply chain security in the automotive industry

The security of the software and hardware supply chain is of paramount importance in cybersecurity. The increased uptake of AI technologies has further amplified this issue with the addition of complex and opaque ML algorithms, dedicated AI modules and third party pre-trained models that now become part of the supply chain. The particularities of the supply chains in the automotive sector, with large and complex dependencies on both hardware and software, add to this complexity.

Proper AI security policies should be established across the supply chain, including third-party providers, ensuring a proper governance and developing an AI security culture across the supply chain. Continuous risk assessment processes supported by threat intelligence could enable the relevant actors to promptly identify and monitor potential AI risks and emerging threats related to the update of AI in autonomous driving. Compliance with specific regulations in the automotive sector (such as UNECE R155 [2]) could be considered to ensure the security of the supply chain.

Cybersecurity processes and controls of AI techniques in autonomous driving

The uptake of AI in autonomous driving brings about important cybersecurity concerns. The increased digitalization of vehicles and the inclusion of AI functionalities result in a larger attack surface and might

significantly increase the incentives for attackers to target AVs. Cyberattacks against AVs do not only concern the particularities related to AI, but also include the security of the underlying digital infrastructure and related digital systems. It is thus crucial to evolve existing security processes and practices to consider this increased uptake of AI technologies and digitalization in vehicles, particularly in the context of autonomous driving.

The automotive industry should embrace a security by design approach for the development and deployment of AI functionalities. This could include the usage of standardised approaches and homogeneous interoperable AI solutions for automotive systems. It is important to promote a culture of cybersecurity (particularly on AI enabled vehicles) across the automotive ecosystem, developing best practices, promoting R&D and innovation and progressively integrating cybersecurity controls and assessments in the current industry processes connected to the lifecycle of autonomous driving AI products and services.

Increase preparedness and incident response capabilities

The current cybersecurity landscape connected to the uptake of AI in AVs is limited to theoretical analysis and experimental use case studies carried out in laboratories and controlled environments. However, the expected increase in the deployment of higher levels of automation in road vehicles could quickly change this picture.

It is important that the automotive sector increases its level of preparedness and reinforces its incident response capabilities to handle emerging cybersecurity issues connected to AI. This includes the establishment of cybersecurity incident handling and response plans based on standards, including vulnerability management processes and patch deployment strategies. Cyber exercises in the form of simulations can also be of help to better understand potential impacts of newly discovered vulnerabilities, raise awareness in the organisations, train the several actors, and evaluate existing plans and procedures.

Increase capacity and expertise on AI cybersecurity for automotive systems

The digital transformation experienced by the automotive sector in the last decade with the growth of the adoption of digital components in vehicles has driven the industry to increasingly face cybersecurity challenges. The uptake of AI as an enabler for autonomous driving vehicles will further amplify this trend placing cybersecurity as one of the critical requirements to ensure safety and promote trust.

In this respect, the lack of sufficient security knowledge and expertise among developers and system designers on AI cybersecurity is a major barrier that hampers the integration of security in the automotive sector. The proper application of the security by design principle requires that all actors involved in the lifecycle of the product are sufficiently proficient on cybersecurity and work systematically together towards the common goal of building a secure product. AI cybersecurity cannot just be an afterthought where security controls are implemented as add-ons and defence strategies are of reactive nature.

Particularly in the automotive sector, cybersecurity is a multidisciplinary endeavour. This is especially true for AI systems that are usually designed by computer scientists and further implemented and integrated by engineers. AI systems should be designed, implemented and deployed by teams where the automotive domain expert, the ML expert and the cybersecurity expert collaborate.

1. INTRODUCTION

Advances in Artificial Intelligence (AI) have opened a whole new realm of opportunities in many areas of our connected digital society. The possibility to automate large parts of our daily activities, considered until now as out of the reach of computing machines, offers new perspectives to address the many challenges humans are facing. In the transportation sector, AI is playing a key role in the development of new generations of cars that will provide autonomous and semi-autonomous driving services to passengers and enable high levels of automation, with tangible benefits in terms of road mortality, traffic congestion, or mobility opportunities.

In this respect, AI is utilised as a means to enhance service provisioning and offer more secure and safe driving conditions. However, at the same time, there are security implications of AI to the entire ecosystem of digital products and services. AI enables new use-cases where cyber impacts cross the barrier between the digital and physical world and can translate into serious safety problems. The automotive sector constitutes a high-risk domain, which is directly affected by the risks associated with these cybersecurity issues. In fact, automotive security is tightly linked to safety: cyberattacks can cause safety problems and harms in the physical world, potentially at large scale. All of this constitutes a reason for focusing efforts on risk alleviation in this sector.

This report aims to provide insights on these cybersecurity challenges, specifically connected to the uptake of AI techniques in autonomous driving vehicles.

1.1 Definitions

The SAE J3016 standard [3] defines six levels of driving automation for on-road vehicles, ranging from level 0 with no driving automation at all to level 5 with full driving automation and no need for a driver, as shown in Figure 1.

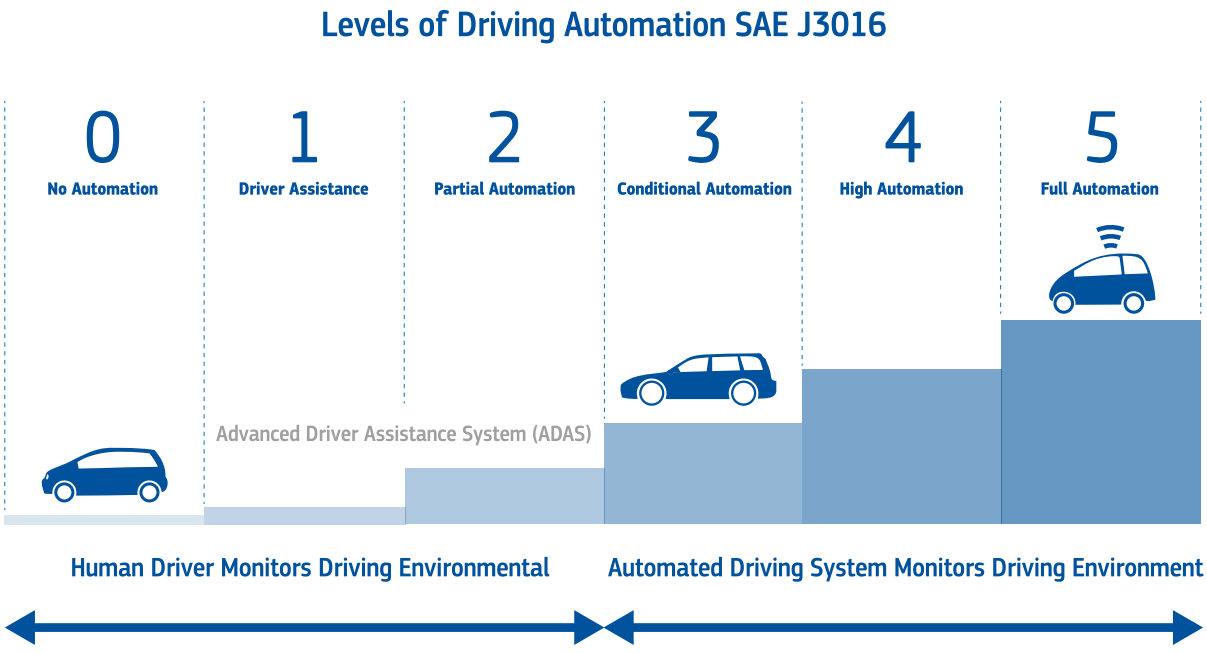

Figure 1. Vehicles automation levels as defined in SAE J3016.

This study focuses on **semi-autonomous and autonomous cars**, which are also referred to as Automated Driving System-Dedicated Vehicle (ADS-DV) in SAE J3016 standard, defined as follows:

- **Semi-autonomous cars (level 4 of automation):** refers to highly automated cars that are equipped with a multitude of sensors in order to be 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).
- **Autonomous cars (level 5 of automation):** refers to fully automated cars that are equipped with a multitude of sensors in order to be able to autonomously perform all driving functions under all conditions (i.e. at any time and on any road). This type of car may not even include a steering wheel or accelerator/brake pedals.

Various national and international bodies have adopted the definition of the SAE standard for autonomous cars, among them the Australian National Transport Commission [4], the Government of Japan [5], the Government of Singapore [6], the UK's Department for Transport (DfT) [7], the US National Highway Traffic Safety Administration (NHTSA) [8], the Government of Ontario, Canada [9] and the European Road Transport Research Advisory Council (ERTRAC) [10].

Besides automating driving, another innovation trend in the car industry consists in providing unprecedented levels of connectivity. Connectivity supports the communication of vehicles with all sorts of infrastructures and devices, and may serve (*i*) increasing functionalities and services for drivers, (*ii*) integrating information needed to enact autonomous driving and (*iii*) enable new driving patterns such as vehicle platooning [11]. Commonly available connection modalities comprise:

- **Vehicle-to-Network (V2N)** connects the vehicle to Internet and/or the cloud, to enable exchange real-time information about traffic, routes, and road situation. This connection is at the base of infotainment systems and an option available in most of the current vehicles.
- **Vehicle-to-Vehicle (V2V)** connects vehicles to exchange information comprising their respective location, direction, speed, braking status, and steering wheel position. Since V2V technology enables sensor outreach of neighbour cars, it may be an enabler of autonomous driving integrating the on-board sensing of the environment.
- **Vehicle-to-Infrastructure (V2I) and Infrastructure-to-Vehicle (I2V)** technologies allow vehicles to communicate with road infrastructure and vice versa to support variety of traffic management applications and services.
- **Vehicle-to-Person (V2P)** technology enables vehicle's connection to smartphones and wearable devices, so that pedestrians or any other vulnerable road user (e.g. cyclists, e-scooter users, etc.) can share data with cars. This may be used to share location information and coordinate the operation of the vehicle with pedestrian's behaviour (e.g. alerting drivers if, for instance, they need more time to cross the road
- **Vehicle-to-Device (V2D)** and Vehicle-to-Everything (V2X) technologies enable the connection of vehicles with any surrounded device, object, and infrastructure connected to the Internet.

The combination of the two trends (toward connected and AVs) will eventually result in the full development of Cooperative, connected and automated mobility (CCAM) [12], in which Connected Autonomous Vehicles (CAVs) are expected to improve significantly road safety, traffic efficiency and comfort of driving, by helping the driver to take the right decisions and adapt in real-time to the traffic situation.

Modern vehicles currently available on the market already include Advanced Driver Assistance Systems (ADAS) as safety features that assist drivers in specific circumstances, such as keeping the car from drifting out of the lane or helping the driver stop in time to avoid a crash or reduce its severity. Advanced functionalities required in such autonomous and connected systems heavily rely on methods for data acquisition, communication, processing and understanding, which empower the vehicle to sense the inner and outer environment and make decisions. AI and its subfield of Machine Learning (ML) are the core enabling technologies of such functionalities. It is easy to foresee that the uptake of these new technologies will introduce new vulnerabilities. Outlining the range of vulnerabilities that vehicles featuring these systems may exhibit is an important goal of this report. To this end, an overview of the functional architecture underlying CCAMs and ADAS may prove useful to frame how and where AI and ML get involved.

AVs have been under development for a relatively long time, and numerous approaches and proofs-of-concept have been developed to provide solutions to the tasks discussed above, however, without reaching a sufficient level of maturity to be implemented as an end-user product. The rise of AI and ML techniques based on Deep Learning (DL) has been a game-changer, with major breakthroughs in computer vision and behaviour modelling that galvanized the industry.

Many companies have developed prototypes that are based on these techniques, and cars with high level of autonomy are already commercialized.

Today, AI and ML are the keystone of highly-accessorised smart cars and will be the key for the next-generation cars, whose driving experience will be more and more ameliorated with user services and automated assistance towards more secure and safe driving. Not by chance, many car manufactures are currently running media campaigns advertising the use of AI in their products [13]. There is a growing tendency in the automotive sector towards prioritizing security in both hardware and software, often in collaboration with academic research centres. The ultimate goal of this escalating use of AI is fully autonomous driving on automation level 5 and the delivery of AVs for both public and private usage.

1.2 Scope

The scope of this report focuses on the analysis of the cybersecurity challenges specifically connected to the uptake of AI techniques in autonomous driving, considering the AI specific cybersecurity issues that surface on top of the more general cyber risks connected to digital systems.

Cybersecurity of digital systems in general, including those supporting AI, lies outside the scope of this work. The interested reader is referred to related work on the topic of securing AI, and in particular the recently published ENISA AI Threat Landscape [1]. The threat landscape serves as a baseline for the identification of relevant assets and threats in the AI ecosystem and was developed collaboratively with the ENISA ad hoc Working Group on Artificial Intelligence Security [14], in which JRC is actively participating.

- The main contributions of this report are summarized below:
- State-of-the-art literature survey on AI in the context of AVs.
- Mapping of AVs' functions to their respective AI techniques.
- Analysis of cybersecurity vulnerabilities of AI in the context of autonomous driving.
- Presentation and illustration (theoretical and experimental) of possible attack scenarios against the AI components of vehicles.
- Presentation of challenges and corresponding recommendations to enhance security of AI in autonomous driving.

1.3 Target audience

This study focuses on the cybersecurity of AI components and systems for AVs. A set of challenges and recommendations is also provided to improve AI security in this field and mitigate potential threats and risks. Hence, the target audience of this report comprises the following profiles:

- **Policy makers** will be informed about the AI particularities in autonomous driving in order to establish proper AI security policies across the automotive supply chain.
- **Regulatory bodies** will better understand the AI security needs in the automotive industry in order to invest on efforts for the development of cybersecurity regulations incorporating the AI special characteristics.
- **Standardisation bodies** will be informed about the AI particularities in autonomous driving in order to strive for standardised AI components/solutions and standards that incorporate AI in AVs aiming to ensure properly secure vehicles.
- **National authorities** will be further informed about the AI cybersecurity in autonomous driving in order to agree on cybersecurity policies that capture the particularities of AI in automotive systems.
- **Original Equipment Manufacturers (OEMs)** will better understand the AI particularities as a first step to incorporate secure AI components while they design new cars and handle the assembly of the various car components.
- **Tier 1 and Tier 2 car components suppliers** will better understand the special characteristics of AI and the need for secure AI solutions in automotive industry, in order to provide car components that incorporate secure AI solutions.
- **AI developers** will be informed about the special needs of security in automotive industry in order to develop AI components and systems having security in mind.

1.4 EU and international policy context

Various attacks in the automotive context [15]–[20], either against AI or not, which were publicly reported over the last three years, led to a relatively quick awareness of policy makers, regulatory bodies and the automotive industry for the security needs and the development of several cybersecurity regulations and initiatives [21], aiming to ensure properly secure vehicles, as presented below.

At the European level:

- Early 2014, the European Commission's Directorate-General for Mobility and Transport (DG MOVE) set up a Cooperative Intelligent Transport Systems (C-ITS) deployment platform. This latter was conceived as a cooperative framework including national authorities, stakeholders and the European Commission, with the objective to identify and agree on how to ensure interoperability of C-ITS across borders and along the whole value chain, as well as to identify the most likely and suitable deployment scenario(s).
- In 2016, the European Commission adopted a European Strategy on Cooperative Intelligent Transport Systems, a milestone initiative towards cooperative, connected and automated mobility [22]. The objective of the C-ITS Strategy is to facilitate the convergence of investments and regulatory frameworks across the EU, in order to see deployment of mature C-ITS services in 2019 and beyond [23].
- In 2016, the Member States and the European Commission launched the C-Roads Platform to link C-ITS deployment activities, jointly develop and share technical specifications and to verify interoperability through cross-site testing. Initially created for C-ITS deployment initiatives co-funded by the EU, C-Roads is open to all deployment activities for interoperability testing.
- In 2017, the European Commission's Directorate-General for Internal Market, Industry, Entrepreneurship and Small and Medium-sized Enterprises (SMEs) (DG GROW) launched an initiative on safety regulations with the aim to contribute to a further decrease of the number of road fatalities and injuries considering amendments to the General Safety Regulation and the Pedestrian Safety Regulation.
- In 2018, the European Commission published the EU Strategy for mobility of the future [24]. This strategy sets out a specific action to implement a pilot on common EU-wide cybersecurity infrastructures and processes that are needed for secure and trustful communication between vehicles and infrastructure for road safety and traffic management. Since 2018 the European Commission is implementing the EU C-ITS Security Credential Management System (EU CCMS) based on the European C-ITS Security Policy (SP) and C-ITS Certificate Policy (CP) published on the website of the C-ITS Point of Contact (CPOC) [25].
- In 2019, the European Commission has set up a Commission Expert group on cooperative, connected, automated and autonomous mobility, named "CCAM" [26], [27], to provide advice and support to the Commission in the field of testing and pre-deployment activities for CCAM. In 2020, to successfully implement the pilot on common EU-wide cybersecurity infrastructures and processes, a sub-group on C-ITS under the Commission Expert Group on Intelligent Transport Systems [12] was set up. The sub-group's task shall be to assist the Commission in working on the implementation of the aforementioned pilot and to foster exchange of experience and good practice in the field.
- In September 2020, the European Commission published a report by an independent group of experts on Ethics of Connected and Automated Vehicles [28]. The report 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.
- The protection of road users' privacy and personal information is also addressed by the recent EU General Data Protection Regulation (GDPR) [29], which officially went into effect in May 2018.
- The Network and Information Security directive (NIS) [30] also addresses AVs' cybersecurity issues as it intends to provide generic security measures in order to enhance cybersecurity across EU.

International Context

- Several international cybersecurity standards and recommendation documents are also under development. In particular, the United Nations Economic Commission for Europe (UNECE) has issued a regulation on cybersecurity [2] which defines a set of requirements that shall be fulfilled by vehicle manufacturers, suppliers and service providers, covering the entire vehicle lifecycle (i.e. from the vehicle development to its decommissioning).
- Transport Canada released in 2020 "Canada's Vehicle Cyber Security Guidance" [31], which provides guiding principles to help ensure vehicles are cyber-safe. This Cyber Guidance aims to support industry stakeholders by providing technology neutral and non-prescriptive guiding principles to strengthen cyber security throughout the vehicle lifecycle.
- The European OEMs published a set of cybersecurity principles, through the ACEA Principles of Automobile Cybersecurity [32], which are already implemented by OEM companies.
- The National Highway Traffic Safety Administration (NHTSA) from the U.S. government issued in late 2016 a document introducing several cybersecurity best practices for smart cars [33].
- The Singapore Standards Council released in 2019 a set of guidelines for the deployment of AVs

called Technical Reference 68 [6], whose Part 3 is related to define cybersecurity principles and an assessment framework.

- China's National Development and Reform Corporation (NDRC) updated in February 2020 its "Strategy for Development of Intelligent Vehicles" [34], which establishes five key missions, including the establishment of a "comprehensive cybersecurity system".
- The US Automotive Information Sharing and Analysis Center (Auto-ISAC) has been maintaining since 2016 a series of Automotive Cybersecurity Best Practices [35], which provide guidance on the implementation of automotive cybersecurity principles.

Standards

- The British Standards Institution (BSI) Group published in December 2018 two Publicly Available Specifications (PAS), namely PAS 1885 [36] and PAS 11281 [37]. The former, which is entitled "The fundamental principles of automotive cyber security", provides high-level guidance to provide and maintain cybersecurity. As regards to PAS 11281, entitled "Connected automotive ecosystems – Impact of security on safety – Code of practice", it provides recommendations for managing security risks in a connected automotive ecosystem.
- The European Telecommunications Standards Institute (ETSI) has been developing a set of technical specifications [38]–[41] to define an Intelligent Transport System (ITS) security architecture along with services specification to ensure information confidentiality and prevent unauthorized access to ITS services. They also address the trust and privacy management for ITS communications. These standards are integral foundation of the European C-ITS Certificate and Security policies [42], which are the governing policy documents enforcing some of the ETSI standards as baseline for interoperable and secure deployment of C-ITS in the EU.
- The standard of Society of Automotive Engineers SAE J3061 [43], officially published in January 2016, is considered as the first standard addressing automotive cybersecurity. It provides a set of high-level cybersecurity principles and guidance for cyber-physical vehicle systems.
- The International Organization for Standardization (ISO) and SAE collaborated to supersede the SAE J3061 recommended practice and proposed the ISO/SAE 21434 [44]. This standard focuses on automotive cybersecurity engineering by specifying requirements and providing recommendations for cybersecurity risk management for cars (including their components, software and interfaces) all along their entire lifecycle. Finally, SAE J3101 [45] defines common requirements for security to be implemented in hardware for ground vehicles.

In 2019, ENISA performed, with the involvement of the JRC, a study on "Good practices for security of smart cars" focused on semi-autonomous and autonomous cars [46]. Moreover, in 2016, ENISA has performed a study on smart cars security issues, which resulted in a document entitled "Cyber Security and Resilience of smart cars" [47]. In 2020, JRC published a report on the future of road transport [234]. In the same year, ENISA established the Connected and Automated Mobility Security (CAMSec) experts Group, to address the cybersecurity threats, challenges and solutions of Intelligent Transport Systems (ITS) and CAM Transport. The members of CAMSec are vehicle manufacturers with focus on cybersecurity, suppliers and developers of embedded hardware/software for smart cars, associations and non-profit organisations involved in vehicle security, road authorities and academia, as well as standardisation bodies and policy makers. In previous years, ENISA has also established the Cars and Roads SECurity (CaRSEC) working group which addresses smart cars cybersecurity threats, challenges and solutions to protect road users' safety. JRC contributes to these security expert groups.

In parallel, the use of AI techniques for decision-making systems in high-risk domains [48], including autonomous driving, has led to a growing awareness of the shortcomings of current AI systems and has raised concerns in society about the compliance of AI systems with respect to a certain number of requirements, including explainability, fairness, reliability or transparency. These recent years, many proposals have been published by public and private actors to establish principles which AI systems should follow to ensure they will respect fundamental rights, and act in a safe and secure manner. The European Commission is particularly active on this topic, with the establishment of multiple initiatives to ensure *trustworthy AI* [49] at the service of the citizens.

2. AI TECHNIQUES IN AUTOMOTIVE FUNCTIONS

2.1 AI in autonomous vehicles

The last decade has seen an increase of efforts towards the development of AVs. An AV is a driving system that observes and understands its environment, makes decisions to safely, smoothly reach a desired location, and takes actions based on these decisions to control the vehicle. A key enabler of this race towards fully AVs are the recent advances in AI, and in particular in ML. Designing an AV is a challenging problem that requires tackling a wide range of environmental conditions (lightning, weather, etc.) and multiple complex tasks such as:

- Road following
- Obstacle avoidance
- Abiding with the legislation
- Smooth driving style
- Manoeuvre coordination with other elements of the ecosystem (e.g. vehicles, scooters, bikes, pedestrians, etc.)
- Control of the commands of the vehicle

Usually, autonomous driving is described as a sequential perception-planning-control pipeline, each of the stages being designed to solve one specific group of tasks [50]. The pipeline considers input data, generally from sensors, and returns commands to the actuators of the vehicle. The main components of a driving-assistant as well as of an AV are broadly grouped into hardware and software components. The hardware component includes sensors, V2X facilities, and actuators for control. The software part comprises methods to implement the vehicle perception, planning, decision and control capability. Figure 3 displays typical elements of this pipeline. They are implemented by decomposing each problem into smaller tasks, and developing independent models, usually using ML, for each of these tasks.

This chapter is structured as follows: First, a brief introduction to the main high-level automotive functions where AI plays an important role is given, as well as a presentation of the main hardware sensors that can be found on vehicles, and that generate the data that are processed by AI software components. After these two sections, a description of the main AI techniques commonly used is given, followed by a discussion on how these techniques are leveraged to implement the high-level functions in AVs. Finally, a summary of the chapter is presented in the form of three tables, highlighting the links between functions, hardware and software components, and techniques.

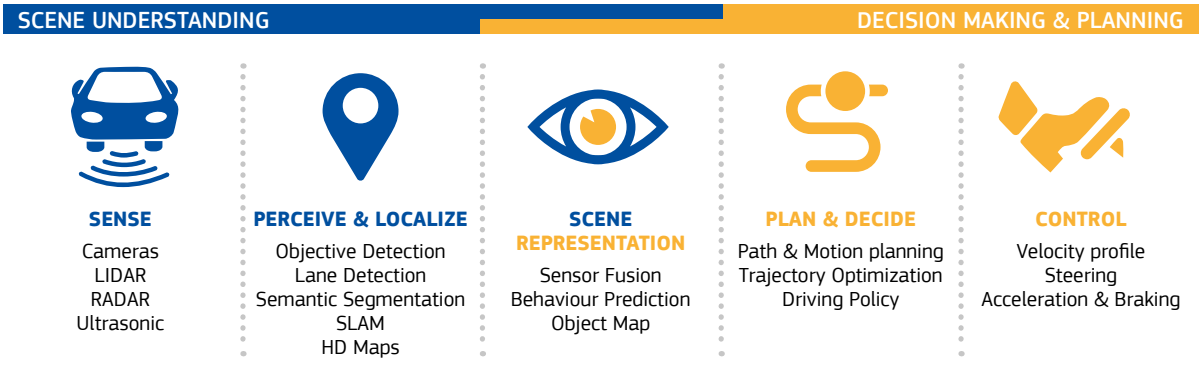

Figure 2. Typical elements of autonomous driving systems. Inputs from the environment are obtained from the sensors of the vehicle or external mapping information. They are used to perceive and understand the environment, plan the trajectory of the vehicle, and act on the vehicle's commands.

2.1.1 High-level automotive functions

Currently, fully autonomous driving solutions are being mostly experimented with prototypes. Nonetheless, vehicles with levels of automation up to level 3 are already on the road, with driving assistance functionalities relying on AI and ML. Technology-enhanced functionalities featured by commercialised vehicles that leverage the use of AI and ML are, for instance, braking assistance, smart parking, or vocal interactions with the infotainment system.

Features of AVs can be decomposed into several high-level automotive functions that are typically used by car manufacturers to advertise the autonomous capabilities of their products. As of today, the technical specifications of such functions are not uniformly defined and vary between manufacturers. In the following, we provide a non-exhaustive list of the most common automotive functions that are deemed as essential to achieve autonomous driving [10]. It is worth noting that most functions have been primarily designed to assist drivers rather than replace them (in vehicles with a level of autonomy from 1 to 3), by providing warnings, or taking control of the vehicles in limited situations. With fully developed AVs, these functions are part of the driving process and, essentially, contribute to replacing the driver1. At the end of this chapter, the following functions are considered and are mapped to specific AI tasks:

- **Adaptive cruise control (ACC)** consists in adjusting the speed of the vehicle in order to maintain an optimal distance from vehicles ahead. ACC estimates the distance between vehicles and accelerate or decelerate to preserve the right distance [51].
- **Automatic Parking** (or parking assistance) systems consist in moving the vehicle from a traffic lane into a car park. This includes taking into account the markings on the road, the surroundings vehicles, and the space available, and generate a sequence of commands to perform the manoeuvre [52].
- **Automotive navigation** consists in finding directions to reach the desired destination, using position data provided by GNSS devices and the position of the vehicle in the perceived environment [53].
- **Blind spot / cross traffic / lane change assistance** consists in the detection of vehicles and pedestrians located on the side, behind and in front of the vehicle, e.g. when the vehicle turns in an intersection or when it changes lanes. Detection is usually performed using sensors located in different points of the car [54], [55].
- **Collision avoidance (or forward collision warning) systems**, consist in detecting potential forward collisions, and monitoring the speed to avoid them. These systems typically estimate the location and the speed of forward vehicles, pedestrians, or objects blocking a road, and react proactively to situations where a collision might happen.
- **Automated lane keeping systems (ALKS)** consist in keeping the vehicle centred in its traffic lane, through steering. This includes the detection of lane markings, the estimation of the trajectory of the lane in possible challenging conditions, and the generation of actions to steer the vehicle [56].
- **Traffic sign recognition** consists in recognizing the traffic signs put on the road and more generally all traffic markings giving driving instructions, such as traffic lights, road markings or signs. This implies to detect from camera sensors various indicators based on shape, colours, symbols, and texts [57].
- **Environmental sound detection:** consists in the detection and interpretation of environmental sounds that are relevant in a driving context, such as horn honking or sirens. This requires performing sound event detection in noisy situations.

In what follows, we first analyse the standard blocks of hardware and sensor components. We then give a brief overview over the most important AI techniques and their software realization used for designing AVs. The chapter concludes by mapping automotive functions to AI functions in order to facilitate the identification of relevant vulnerabilities and cybersecurity threats in autonomous driving. By narrowing down the AI techniques that are actually used in AVs, one scopes down the problem of identifying pertinent cybersecurity threats related to the use of AI in autonomous driving.

2.2 Hardware and sensors

Humans drive cars by taking actions with hands and feet, based on decisions made considering the input received from our senses, mainly sight and hearing. Similarly, AVs rely on a variety of sensors to observe the surroundings and provide data to the AI systems of the vehicle, and on actuators to control the motion of the vehicle. The hardware components allow

1 To that end, we do not consider these functions as warning systems, nor did we include functions that are necessarily acting on the behaviour of the driver (such as driver drowsiness detection). Likewise, functions that increase safety but are not directly linked to cognitive capabilities, such as anti-lock braking system (ABS) or tire pressure monitoring, are not mentioned.

the vehicle to sense the outside surroundings as well as the inside environment via specific sensors, to act via the actuators that regulate the car movement, and to communicate with other agents/devices via the V2X technology.

Sensors, as the primary source of information for AI systems, are a critical element of AVs. All sensors can be broadly classified in three distinct groups [58]:

- Exteroceptive sensors are those sensors that are designed to perceive the environment that surrounds the vehicle. They are relatively new sensors present in cars, and are the eyes and ears of the car. Cameras and Light Detection and Ranging system (LIDARs) are the main vectors of information for driving purposes. Other sensors, such as Global Navigation Satellite Systems (GNSSs), Inertial Measurement Unit (IMU), radars and ultrasonic sensors, are also used to probe the environment, but tend to be limited to specific tasks (e.g. close detections, sound listening) or to add redundancy, increasing the reliability of results in the case of malfunction of a sensor.
- Proprioceptive sensors, on the other hand, are those that take measurements within the vehicle itself. They have been present in cars for decades, and are mostly used for control purposes. They include the set of analogue measurements that are encoded in digital form indicating values such as the engine's revolution per minute (RPM), speed of the car (as measured by wheel's rotation), direction of steering wheel, etc.
- Other sensors are those sending the information that the vehicle might receive from its digital communication with other vehicles, V2V communications or V2I. They mainly concern the connected infrastructure of vehicles, and therefore they will not be discussed in the rest of the report.

The integration of sensors in vehicles varies according to carmakers [59], [60] and depends on the software strategy chosen to process the streams of data. Very often, the inputs from multiple sensors are combined in a process called data fusion [61] to align all data streams before processing, as sensors are usually providing images from different natures (2D images, 3D point clouds, etc.) with different temporal and spatial resolution.

Table 1 presents the main characteristics of the most common sensors found on autonomous cars, in addition to the LIDARs and cameras. The localization of these sensors on the vehicle and their main uses are illustrated in Figure 3.

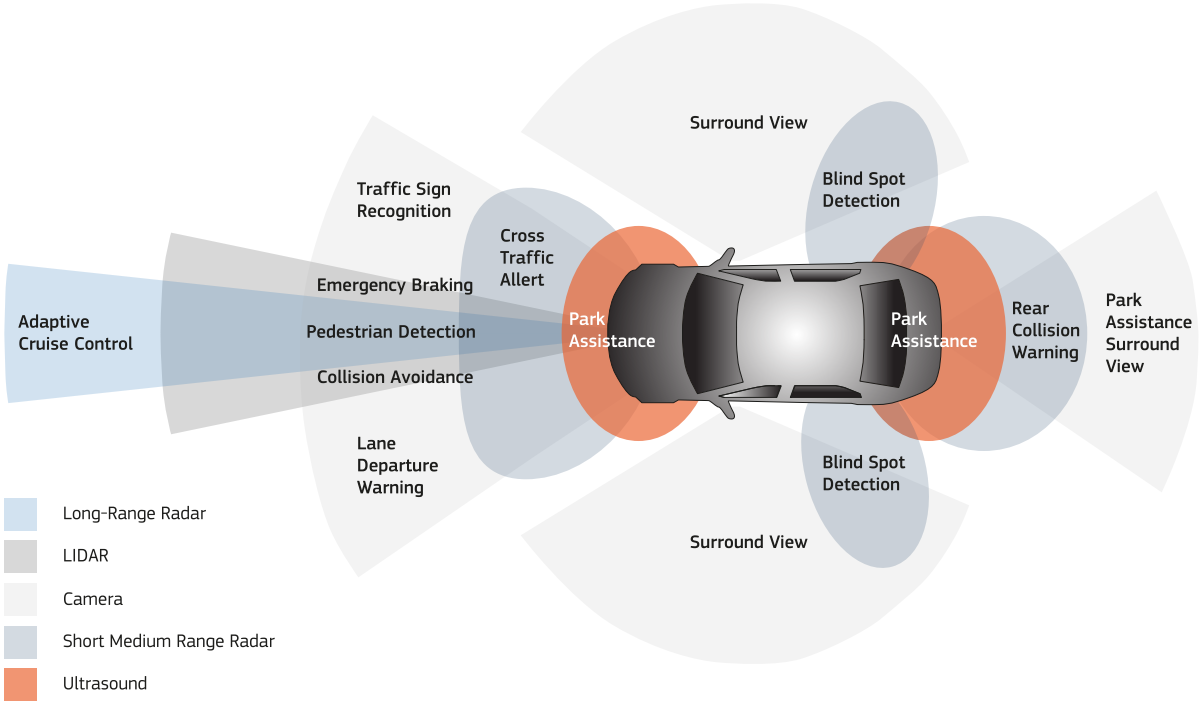

Figure 5. Localization of the sensors on the vehicle and their main uses.

2.2.1 LIDARs and cameras for computer vision

Cameras and LiDARs are the most widespread sensors in autonomous cars, used to reproduce and enhance human vision. Digital video cameras are able to obtain a 2D representation of the 3-dimensional world. They provide a stream (video feed, as a sequence of images) of 2D maps of points (pixels) encoding colour information. Computer stereo vision techniques can be applied using multiple cameras and/or considering the different images in relation to the known movement of the vehicle. Examples of images from cameras are depicted in Figure 4.

A LiDAR illuminates the environments with lasers and collects the reflected light. The analysis of the signal received allows the generation of a depth map of the scene (see Figure 4). The depth map is further processed to recreate 3D maps of the environment [54] considering missing values in the acquired 3D data points, unexpected reflections due to wrong perception of surfaces, and many other issues that may appear during the acquisition in real world scenarios.

Compared to LiDARs, cameras have the advantage that they distinguish colours, allowing the recognition of elements such as road signs, traffic lights, vehicle lights or text warnings. However, cameras also exhibit certain limitations compared to LiDAR: camera vision could be impaired by certain weather conditions such as rain, fog or sudden light changes such as when a vehicle gets out of a tunnel, while these conditions would affect to a lesser degree a LiDAR system.

	Sensor Type	Range	Pros	Cons
Exteroceptive (sensors that perceive environment)	LiDAR	Up to 200 meters	High precision High accuracy Wide Field of View	High cost No colour information Worsen aerodynamics (usually mounted on the roof)
	Cameras	Up to 100 meters	Can see colours and textures Low cost High availability	Sensitive to low intensity light Heavily affected by adverse weather conditions Inaccurate range estimation
	Radar	5 meters – 200 meters	Robust to environmental conditions Cheaper than LiDAR Mature and readily available Capable of determining relative motion of objects Fast detection response	Noisy response for metallic objects Not suitable for static objects Poor lateral resolution
	Sound microphone	Several hundreds of meters	Allows to hear environmental sounds.	Limited to audio signals.
	Ultrasonic sensors	Up to 2 meters	Robust to adverse weather conditions Proven track of reliability Most accurate sensor for close proximity Inexpensive	Only suitable for very short range Low resolution Not suitable for high speeds Heavily affected by changes in environmental conditions (temperature, humidity)
Proprioceptive (sensors that measure values within the system)	GNSS		High accuracy. Relatively inexpensive. Widespread deployment High-integrity and high-precision positioning capabilities	GNSS signals do not penetrate buildings such as multi-story car parks or inside tunnels, Issues of reflectivity and satellite visibility in built-up urban areas. Vulnerability to intentional and unintentional signal interferences.
	IMU	Within the vehicle	Needs no connection to or knowledge of the external world 6 degrees of freedom Used in sensor fusion with other localization techniques Inexpensive	Accuracy is dependent on calibration of accelerometer and three axis rate sensor. Around 30cm accuracy, so needs to be used in combination with other sensors
	Encoders (position, velocity, etc.)	Within the vehicle	Gives an accurate state of the vehicle Low cost. Easy to install.	Limited accuracy.

Table 1. Comparison of AV sensors [58], [62].

Figure 4. Examples of images from an online set of Italian traffic signs [63] captured by a camera in three different environmental conditions (top) daytime (middle) fog (bottom) night-time.

Figure 5. Superposition of outputs from cameras (RGB images) and from LIDARs (range maps) (adapted from the Waymo Open Dataset [64]). For the range, the colour is coded from yellow (close) to purple (far).

2.3 AI techniques

AI is generally defined as a collection of methods capable of rational and autonomous reasoning, action or decision making, adaptation to complex environments and/or to previously unseen circumstances [65]. AI was initially born as an academic discipline in the second half of the twentieth century and led since then to significant advances in the automation of some human level tasks, nonetheless without much impact beyond academic circles for a long time [66]. It is deeply rooted in the fields of computer science, discrete mathematics and statistics, with an eventful history before gaining the popularity that makes it nowadays a key domain of the current digital revolution, thanks to the tremendous performances achieved by modern systems.

Typical problems related to AI require the development of programs able to demonstrate some forms of reasoning, knowledge representation, planning, learning, and, more generally, cognitive capabilities. These competencies are usually considered as being natural to humans but are difficult to translate explicitly into algorithms. Nowadays, from a scientific perspective, AI is actually a heterogeneous field that regroups different subfields with diverse views on how to address these problems.

Research on AVs was historically pioneered in the field of robotics, with several cars in the 1980s, and later on, able to drive autonomously in controlled environments. Nonetheless, the complexity of real-world environments, and the necessary reliability that are required for such vehicles, has curbed their development until the significant recent progress made in ML. Since the last decade, tremendous milestones have been reached in computer vision, natural language processing or game reasoning, pushing autonomous driving a leap forward. Although some functions are still solved using traditional methods, ML is increasingly used, relying on the huge quantity of data that are collected by companies, with millions of kilometres travelled by autonomous cars

under human supervision in real-world conditions or using simulated environments.

In the following, a short introduction to the relevant fields of ML for autonomous driving is provided, Besides giving the technical basis that will support the discussion in the rest of the report, the goal of this section is also to highlight the diversity of techniques that are being employed for the different tasks, and the complexity of the full processing chain.

2.3.1 Machine learning: paradigms and methodologies

ML is the scientific field dedicated to the study of models that are able to improve automatically through experience [67]. This acquisition of experience can take different forms, and is usually achieved by extracting relevant patterns from large collection of data. Machine learning algorithms are therefore able to achieve high performance for a variety of complex tasks, hard to solve using conventional programming techniques, without being explicitly instructed how to perform them. Prominent examples include recognizing faces in a picture, identifying objects in video streams, predicting the price of an asset quoted in a financial market, grouping users on an online platform based on their activities, recognizing the emotion of a person, or teaching a robot to move in an unknown environment.

The central element of ML systems is the model that takes as inputs a set of pre-processed data, and returns a prediction. This model is usually described as a mathematical function, with a collection of parameters that have a direct influence on the mapping between the inputs and the predictions. To adapt the model to the desired task, a training stage is performed, and consists in running an algorithm that will update these parameters to fit a training dataset, i.e. a list of samples serving as examples to guide the model towards the expected function. The capability to perform well on data outside the training data, called the generalization, is a desirable property of the resulting model, which is often measured by metrics such as accuracy or mean squared deviation on previously unseen data. Training a model implies applying a host of ad-hoc procedures to increase the generalization capabilities of systems. Popular techniques include data augmentation that consists in artificially increasing the amount of training data by applying random transformations on the training data, and hyperparameter optimization search that tests various settings for the training procedure. The full pipeline includes several additional steps during training and testing that are not detailed here.

2.3.1.1 Paradigms of machine learning

Three different paradigms are commonly considered in ML:

- **Supervised learning** makes use of large and representative set of labelled data to train the model. The underlying problem consists then to return the right label for the input data. Supervised learning includes classification, when the label is discrete (e.g. the make of a car), and regression, when the label is continuous (e.g. the speed of the car). The availability of labelled data is a limiting factor for supervised learning, as labelling can be, in some contexts, expensive and time-consuming.
- **Unsupervised learning** (or self-learning) consists in extracting meaningful patterns from the data without labels by reducing the natural variability of the data, while preserving the similarity or absence of similarity between examples. Unsupervised learning is used for various purposes, such as clustering the samples (e.g. grouping individuals based on their habits) or anomaly detection (e.g. detecting a vehicle with an unusual behaviour).
- **Reinforcement learning** regroups a set of techniques to make models learn sequences of actions in a possibly uncontrolled and/or unknown environment. Contrary to the supervised learning setting, in which the ground-truth is given as labels, learning is guided by indications on how good an action is, given the state of the environment. Consequently, the learning process is dynamic with respect to the feedback it gets from the environment, in a trial-and-error approach.

2.3.1.2 Classical machine learning

A wide range of techniques have laid out the foundation of the field of ML coming from statistics and expert systems, such as linear regression, support vector machine (SVM), k-nearest neighbour (kNN) classifiers, or decision trees. The common point of these methods is that they usually operate on handcrafted features, whose quality can drastically change the performances of the model. Although these techniques show limitations in complex problems such as the ones encountered in computer vision or natural language processing, they are still very popular to solve a large range of problems, in particular when the volume of data is small, when the time available for model training is limited, or if the context domain is well understood.

2.3.1.3 Deep learning

Although the ideas behind neural networks are as old as the field of ML, DL techniques have disrupted the ML landscape these last years, and exported the whole field of AI outside academic circles, thanks to simultaneous progress in computing capabilities, data acquisition and storage, and ML algorithms. The advances in hardware and the digitalisation of the society have permitted to train models on high performance computing infrastructure and to collect huge datasets to do so, in addition of progress made to speed up training algorithms.

One strength of DL is its ability to learn from raw data the most adapted representation for the considered problem, removing the need to handcraft features. DL techniques employ neural networks in layered architectures, denoted deep neural networks (DNN), allowing for flexible designs able to represent relationships between inputs and outputs. Each layer is composed of a number of units called neurons that perform simple linear combinations between the outputs of the previous layer. These stacked architectures exhibits a specialization of groups of neurons in the deepest layers, able to extract more and more complex patterns.

Although powerful, DL is not a silver bullet as it suffers from several limitations that make it impractical in some situations. First, the training of neural network models needs substantial amounts of good quality data and of computational power to be efficient. Secondly, the development of such models, in particular during the training phase, lies on strong engineering practices with limited theoretical guarantees on the overall performances. This severely hinders the understanding of the behaviour of DL models, and is a reason of their vulnerabilities. Thirdly, DNNs are notoriously known to provide accurate results but with an inherent lack of interpretability, making them acting as black boxes. The robustness of such systems with respect to unusual inputs or malicious actions is also under scrutiny by the research community.

2.3.2 Relevant application fields in autonomous driving

2.3.2.1 Computer vision

Computer vision is an interdisciplinary field, at the intersection of ML, robotics, and signal processing, concerned with extracting information from digital images and videos. This covers all stages of the processing chain, from the acquisition of images to the processing and analysis of the image, to the representation of knowledge as numerical or symbolic information. To date, computer vision is the most relevant field of ML with existing applications in AVs. As such, the most significant and well-known vulnerabilities and possible attack scenarios on AI models employed in AVs are involving computer vision techniques. A more detailed focus is then given with respect to other application fields of ML.

Images can take several forms, depending on the type of hardware sensors that have been used to obtain them. Computer vision has been historically interested in the handling of standard RGB images, that represent a large proportion of applications in robotics and image processing, but has also gained more and more interest in the analysis of other forms of images such as 3D point clouds, hyperspectral images, acoustic images, to name but a few. This trend has been significantly fostered with the recent availability of large data sets. In addition to the mode of acquisition, other variables such as the size, the resolution, the quality obtained, the environment of acquisition, etc. have led to specialized sub-domains adapted to specific tasks.

The advent of highly performant convolutional neural networks (CNN) has been a major breakthrough that has drastically changed the technical landscape in computer vision. CNNs are an evolution of DNNs specifically designed to take into account the spatial structure of images [68] by grouping the weights that are locally close. They compensate for one of the drawbacks of fully-connected networks by significantly reducing the number of parameters to learn, making learning on high-dimensional data, which is typically the case of images that are composed of millions of pixels, a more tractable problem. Convolutional layers act as a series of filters that are applied on a small portion of the image to detect a specific pattern such as edges, a specific shape, a dark area, etc., the particularity being that these filters are learned from the data. The size of the filters determines the complexity patterns, and has to be calibrated according the characteristics of the image. CNNs have been successfully used to extract, directly from the raw inputs, efficient representations that are adapted to the problem, and that take into account natural invariances that often appear in images, such as symmetries. Figure 4 illustrates a typical CNN architecture used for classification, and the basic mechanisms at play during the processing.

Today, the overwhelming majority of computer vision techniques are relying in one way or another on

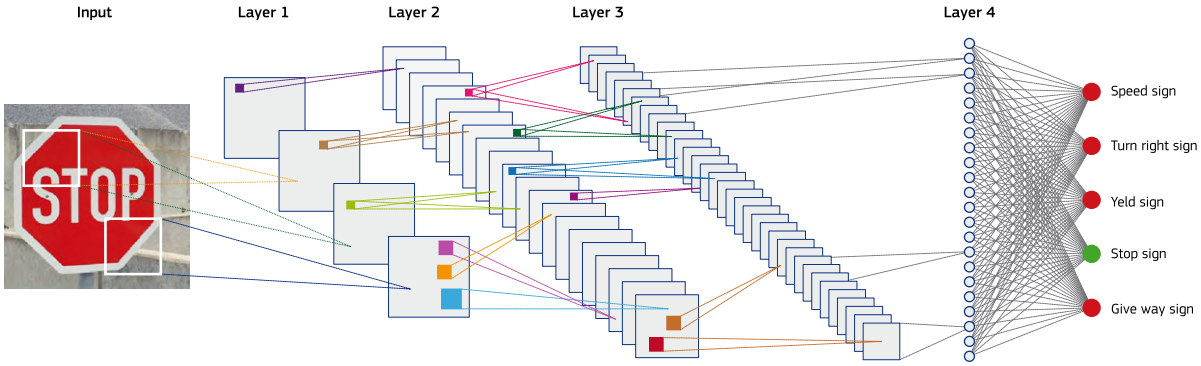

Figure 6. Example of a typical CNN architecture used for classification. Convolutional layers are filters applied on portions of the images. At each layer, the number of intermediate images increases, while their dimensions is reduced. Only a small proportion of links between layers is displayed. The final layer condenses the values to return scores for each class, the highest score being the predicted class.

CNNs, and more and more sophisticated approaches are considered to address always more complicated problems. In the following, the most relevant problems for autonomous driving are briefly summarized.

Object recognition, in its most common form, includes two tasks: detecting and classifying objects in an image. Localization is usually achieved by assigning bounding boxes to regions of the image, while classification assigns to these regions a label from a list of pre-defined categories.

To achieve high performance in the classification task, several architectures have been built by the ML community during the last decade, employing various innovations aimed at increasing the expressive power of models, while limiting at the same time the cost of training. Among them, we can cite: AlexNet [69] is widely considered as the first breakthrough using CNNs; VGG [70] introduces the use of numerous layers, with different size of filters; GoogLeNet [71], [72] makes use of the so-called Inception module, including at the same level various filters of different sizes; ResNet [73] uses shortcut connections between layers to limit the tendency of large models to memorize the training data (also called overfitting); SqueezeNet [74] is designed to be embedded in systems with low capabilities by reducing the number of parameters.

On top of classification architectures, object detection is implemented in two competing designs: single staged and double staged. Double staged approaches split the detection procedure into two stages, region proposal and bounding box search. By far the most widely accepted state-of-the-art double stage design is the family of Fast Region-based NN (R-CNN) [75] architectures. Conversely, single stage design detectors only employ one single network architecture to classify directly pixels or regions. The "You Only Look Once" (YOLO) [76] architecture is an example of a successful single-stage detector, largely used in many recognition systems.

Object recognition in 3D representations, such as point clouds, is becoming more popular these last years, especially with advances in autonomous driving using LiDAR and radar. Albeit less mature than its counterpart in 2D, some first standard architectures have already emerged. Among those are techniques to project the 3D point clouds into a 2D space to use one of the known 2D detectors, for example the YOLO3D architecture [77], or, those techniques that directly develop algorithms working on the 3D data, such as VoxelNet [78] or PointNet [79].

Segmentation is an extension of object recognition that does not consider bounding boxes to delimit objects, but pixel-wise regions acting as masks over the image. A label is then assigned to each region to classify them into prescribed categories. Several types of segmentation problems have been discussed, among them, instance segmentation detects pixels of each object and assigns an identifier for each object; as for semantic segmentation, non-object characteristics, such as sky, water, horizon or textures are part of the elements to segment. In the latter case, the full image is completely segmented, resulting in a label assigned to each individual pixels, forming continuous regions. As for object recognition models, segmentation makes an intensive use of CNNs. State-of-the-art approaches include SegNet [80], EnET [81], PSPNet [82], DeepLab [83], or Mask R-CNN [84].

Vehicle localization, often called Visual Odometry (VO), is a technique to estimate using a sequence of images captured over time by the camera mounted on the vehicle the pose of the camera, i.e. its position and its orientation. A common approach consists in tracking key-point features of clear landmarks and reconstructing the complete pose from these features. Classical VO techniques still dominate the field for AV, although there have been increasingly promising results based on DNNs [85]. DL has also been built on top of classical algorithms implementing outlier rejection scheme in order to discriminate ephemeral from static parts in images [86]. While most techniques focus on 2D images, various proposals have been published to tackle the 3D pose estimation using DL [87].

Tracking of objects is used to determine the dynamics of moving objects. This can be seen as an additional layer on top of image recognition systems, providing for each frame of a video stream the objects present in a scene, the temporal connections between these objects, and a prediction of their future positions. Tracking systems, also referred as MOT (for Multiple Object Tracking) provide this functionality by estimating the heading and velocity of objects, and applying a motion model to predict the trajectory. Tracking is a very active field of research in the computer vision community, and has been studied for a wide range of applications and contexts [88]. Techniques vary according to parameters such as the quality of the detection of objects, the type of data considered, the frame rate, or the nature of the motions involved.

Typically, tracking is solved by assigning an identifier to objects and trying to keep this identifier consistent through successive image frames. This consistence is implemented either by using measures of similarity applied on handcrafted features based on image characteristics, such as colours or gradients [89], or by using CNNs [90]. The modelling of the dynamics is then done using sequential modelling tools to take into account the temporal dependencies between frames. The trend towards models that jointly address multiple tasks concerns also the tracking and the prediction of objects, that is often coupled with object recognition systems. For example, the "Fast and Furious" [91] architecture considers simultaneously 3D detection of objects, their tracking over time, and the forecasting of their motions.

2.3.2.2 Sequence modelling

Modelling sequential data and training predictive systems is a very important subfield of ML with high relevance for autonomous driving. Sequential data encompass data sets that result from dynamical or ordered process introducing a clear sequence and correlation between instances of the data set. Examples include time series modelling, prediction of trajectories, speech and language processing. Sequence modelling plays a significant role in prediction and planning tasks and is used in robotics and signal processing in applications concerned with interpreting continuous flows of environmental data. Classically, the field is dominated by Markov models, autoregressive modelling and dynamical linear filter systems [92], which are built around the assumption of a certain correlation length between successive elements of the series paired with a probabilistic process to model the next element.

Recently, DL and unsupervised representation learning have introduced major advances into the field, mostly in form of specialized network structures, such as recurrent neural networks (RNN) [93] and modern variants such as the Long Short-Term Memory networks (LSTM) [94], able to deal with the sequential nature of data. Key advantages from DL based systems are their capability to easily discover long and short term correlations and to automatically learn representations of dynamical processes, even in complex contexts.

2.3.2.3 Automated planning

Automated planning [95] is a rich field connected to ML at the intersection of other fields such as robotics, complex infrastructure management, decision theory, and probabilistic modelling. It is mainly concerned with the search of optimal strategies, often described as a sequence of actions that should be followed by agents evolving in complex environments, and how to perform them. There exists a wide variety of methods to find the optimal strategy in the specific context of problems that this field aims to address. In the following, we provide a brief description of methods that have been used in an automotive context.

Graph-based planning is used when systems can be represented as networks, including a wide range of applications as diverse as social relationships, transportation or telecommunication networks [96]. In its simplest form, a graph is composed of nodes, that represent the entities, and of edges, that represent a link between the entities. For planning purposes, algorithmic approaches have been used to find optimal trajectories along the edges of the graph to go from one node to another one, using pre-defined constraints. Classical algorithms coming from graph

theory, such as the Dijkstra, Bellman-Ford, or Floyd algorithms [97], are popular approaches that do not require ML techniques to provide satisfactory results.

Deep Reinforcement learning provides a range of methods well suited for planning tasks [98]. It involves learning mapping situations to actions so as to maximize a numerical reward signal. In an essential way, they are closed-loop problems where the learner is not told which actions to take, as in many forms of ML, but instead discovers which actions yield the most reward by trying them out. In the most interesting and challenging cases, actions may affect not only the immediate reward but also the state of the environment and, through that, all subsequent rewards. Many approaches have been developed to find the policy that is optimal, i.e. the policy that returns in average the highest future reward. The use of DL architectures [99] to model the environment has enabled significant advances, with techniques such as deep Q-learning or actor critic learning capable of scaling to previously unsolvable problems.

2.4 AI software in automotive systems

Driving in real world environments with other human-operated vehicles is far from an easy task, which requires complex socio-ethical and decision capabilities able to cope with unexpected and dangerous situations. AI software components embedded in AVs are in charge of reproducing these capabilities, by processing data gathered via the sensors and interpret them in order to decide the action to undertake (e.g. move, stop, slow down, etc.). Three main types of data processing capabilities are involved:

- The **perception** module is responsible to collect multiple streams of data obtained from the sensors, and extract from them relevant information about the environment. This includes the contextual understanding of the scene: detection and tracking over time of vehicles (cars, trucks, bikes, etc.), pedestrians, and objects, and their tracking over time, recognition of traffic signs, traffic lights, marking, lanes, and more generally any element of interest for the driving. The perception module keeps also track of the localization of the vehicle in this environment, detecting its position and orientation with respect to the road and other agents involved in the scene.
- The **planning** module is in charge of the calculation of the trajectory that the vehicle will undertake, considering the route between the start location and the desired destination, as well as all the constraints the vehicle has to respect along the entire path. These constraints include the design of a safe and smooth route taking into account all possible obstacles (still objects, moving vehicles, etc.) and the compliance with driving rules, but also require to take into consideration behavioural aspects due to the presence of humans in the environment.
- The **control** module is responsible to execute the sequences of actions planned by the system by acting on the actuators (speed, steering angle, lights, etc.) to ensure that the trajectory is correctly performed.

The decomposition of the general driving activity into subtasks is a standard approach to address complex problems that enable constructors to select the right methodology for each of the subcomponents. ML has been mostly used in the tasks related to perception, to sense the surroundings and provide a useful representation of the environment. This was spurred by recent advances in computer vision that created industrial opportunities in this sector. Consequently, companies have started to invest in vehicles to collect recordings of driving situations, build up infrastructures to collect, label and process those large datasets, and hire computer vision high-level engineering teams to develop model to address the related tasks. Although solving these problems is still an active area of development, commercial products and services are already in use in modern cars to attain low to medium levels of automation. Conversely, the use of ML for planning and control purposes is still in its infancy, while improvement is going fast, supported by large investments from tech companies. In this context, behaviour modelling techniques are leveraged to learn relevant driving policies that will determine the action to undertake to achieve the trip according to the environment encountered by the vehicle. These problems are different from perception problems, and are still considered as frontier research in the academic community.

Current systems are already achieving tremendous performances in a wide range of conditions, but their capacity to generalize is limited by the extreme complexity and diversity of the world. A crucial challenge still unsolved for ML systems is the right general handling of edge cases, where an unknown situation outside of the training data distribution is encountered. It is very challenging to guarantee that an AI system will output the right results in unusual conditions, leading to potentially hazardous situations, for instance ignoring a stop sign partially covered by snow, or stopping in front of a bush slightly overhanging the side of the road.

In the rest of the section, an overview of the different tasks, and the main techniques that are employed to address them, is presented. This overview is primarily based on works published by the research community and may not reflect accurately the technology embedded in actual semi-autonomous cars, as only limited information is released by car manufacturers on this matter. In addition, perception tasks are prevalently discussed over others, as they constitute currently the main sources of vulnerabilities of AI systems.

2.4.1 Perception

The perception system refers to the ability of an AV to make sense of the raw information coming in through its sensors. It aims at the creation of an intermediate level representation of the environmental state around the vehicle, and at the tracking of the evolution of this state over time. This includes, amongst others, the capability to detect, classify and identify everything that an AV potentially could encounter or has to interact with, such as the infrastructure (roads, signs, traffic lights, etc.), agents (cars, cyclists, pedestrians, etc.), or obstacles. It also consists in the construction of an internal map of the environment, allowing the vehicle to localize itself and other objects of agents in space and time.

The most relevant perception tasks for AVs can be ordered along the terms of scene understanding, scene flow estimation and scene representation and localization. Nowadays, much of the field is dominated by DL techniques, albeit strong influences are coming from robotics, especially for localization and mapping techniques [100], and classical time series pattern recognition filters [101].

2.4.1.1 Scene understanding

Scene understanding encompasses all tasks that aim to provide a current picture of the immediate environment of the AV. Typical tasks include the detection and recognition of all elements present in the environment. Most approaches that fall under scene understanding make use of data streams from various sensors and employ very successful computer vision based architectures. However, understanding objects in a realistic traffic environment in real time poses a number of additional complexities to the theoretically often extremely accurate computer vision systems. This requires indeed taking into account the variability of the scene:

- Variability of appearance: elements present in the environment can have a wide diversity of aspects: objects, actors and surfaces can have various shapes, colours, texture, orientation, brightness, etc.;
- Variability of environments: the environment itself varies according to various factors, including the time of the day, the season, the weather, but also societal factors (maintenance works, strikes, behaviours of agents, etc.);
- Variability of meaning: objects can have different meanings in different contexts or at different times (e.g. traffic signs with time or space constraints), or wrong appearances, for instance in case of reflections.

Identification of roads and lanes This requires distinguishing drivable areas (roads, driveway, trail, etc.) from non-drivable areas (pavement, roundabouts, etc.), the different types of surface, and the various lanes present on the road indicating the direction of the traffic flow. This task has been widely investigated over decades, and has been integrated in vehicles through functions such as lane keeping assistance or lane change assistance. Therefore, such technologies generally do not rely on recent trends in ML, but rather on a wide collection of techniques based on handcrafted features [102] that have proven to be very efficient and reliable. This is done on 2D and 3D images, and is often complemented by real street maps. Despite this, traditional methods tend to be limited to understand the challenging semantics conveyed by lanes on the road, and this task is getting more and more integrated in end-to-end DL systems. This is for instance the case of the Tesla's Autopilot [103], whose the ADAS to detect stop lines along the road is implemented using a DL architecture.

Detection of moving agents and obstacles The detection of moving agents (pedestrians, cyclist, vehicles, etc.) and obstacles (plants, objects, etc.) is mainly addressed using object detection, segmentation, and tracking techniques. Detection and classification of objects from 2D or 3D images such as camera data can be successfully tackled with computer vision architectures, providing that the training dataset is rich enough to characterize fully the diversity of environments. Recognition and segmentation techniques are used to detect drivable areas, objects, pedestrian paths or buildings.

Traffic signs and markings recognition The detection and recognition of the given sign and/or the written indications is crucial to ensure a safe driving. Driving instructions are typically provided according to several vectors:

- **Traffic signs**, usually using a symbolic representation, are the most common driving indicators. While signs vary depending on the shape, the colour, and the pictogram drawn on it, international conventions have helped to achieve a degree of uniformity, despite small local variations. While image-processing approaches have been mainly taking advantage of the well-defined shape and colours of signs, DL has greatly improved the rate of detection [104]. Models based on CNNs are now regularly employed [105], taking advantage of various datasets released for this task [106].
- **Traffic lights** are using a position (usually top, middle, bottom) and colour (usually red, orange, red) code to indicate if the vehicle should stop, prepare to stop, or is allowed to pass. Particular cases should also be taken into account, for instance when a light is blinking. Research works are currently mostly limited by the scarcity of representative public datasets, but are nonetheless led to the development of DL systems including special processing in the colour space [107], [108]. Industrial actors have nonetheless already included traffic light recognition in their vehicles as an assistance feature [109].
- **Textual indications** are also an important way to convey information, in particular in situations where unanticipated rules should apply (e.g. detours, accidents, traffic jam, etc.). Text can be found on traffic signs, painted on the road, or on variable-message signs. Understanding textual indications commonly requires three steps: 1) the detection of the text in the image, 2) the recognition of the characters, 3) the understanding of the meaning. The last step is all the more important since numerous text signs without connection with traffic indications can be found alongside roads, like advertisements or touristic information. Currently, this task is mostly done in an ad-hoc manner, without learning involved. Text detection and recognition in natural scenes have nonetheless been an important area of research, either on static images [110] or videos [111], taking into account artefacts, such as distortions or out-of-focus texts. State-of-the-art approaches have combined convolutional and recurrent neural networks [112], to achieve text recognition and understanding. The availability of datasets [113], [114], although limited, is expected to foster the scientific community to advance research on this ongoing topic that will also benefit autonomous driving.

Previously based on handcrafted features describing images based for example on the colour or the shape, object recognition and semantic segmentation techniques are now systematically used. These techniques require nonetheless adjustments to adapt to the relatively small size of signs and markings compared to bigger objects such as vehicles, as CNNs are typically compressing the image, resulting to small objects, such as signs, to be overlooked. Recognition is also greatly affected by unusual environmental conditions, with degraded performances when the symbols are partially occluded by obstacles or stickers, or hard to distinguish due too direct sunlight or heavy precipitation.

Sound event classification Recognizing environmental sounds is an important aspect for the understanding of the driving scene: many elements, such as tire screeching, honking, or even engine throbbing, convey information about the vicinity of the vehicle, and can help anticipate hazardous situations. This is particularly true for sirens of emergency vehicles that indicate a situation where driving rules have to be adapted. Albeit visual lights are usually present, they may not be visible to the vehicle, for instance in the case of a busy intersection, and ignoring the sound alarm could have dramatic consequences. Sounds also complement the other sensors in low visibility situations.

As for images, sound processing largely makes use of DL techniques [115], either on raw data or on spectrograms (frequency representation of sound signals), even if traditional approaches are still widely used due to the long-standing work on handcrafted features relying on physical and cognitive principles. Related to AVs, few works [116] have been proposed, partly because of the small interest of the AV community on these topics compared to vision, and a lack of availability of dedicated datasets. Detecting relevant sounds in urban areas, especially in noisy situations, is nonetheless an open challenge that will play an important role in the capacity of AVs to achieve human-like performances.

2.4.1.2 Scene flow estimation

Scene flow estimation collects those perception tasks, which are concerned with the dynamical behaviour of the scene, mostly the movement of the objects and vehicles.

Tracking and prediction of moving agents and obstacles

The most important task in scene flow understanding is to track objects and vehicles to predict their individual motion and may require modelling the behaviour of other traffic participants, which is very relevant for planning tasks. The two main challenges

are the tracking of self-motion or stationary objects and the prediction of object motion and behaviour.

Tracking and predicting the motion of objects located in the immediate environment of the AVs rely evidently first on the ability to detect those objects, but pose a number of additional challenges, which led to the development of a class of tracking and prediction methods [117]. The basic object detection problem is extended through a time axis dimension, where individual objects need to be tracked frame by frame, or, conversely their likely trajectory is to be extrapolated into the future. The employed techniques usually rely on elements from sequence modelling, such as probabilistic Markov models or, increasingly, deep recurrent neural network architectures. Both, tracking and prediction – as object detection itself – are conducted using either 2D camera image data, 3D point cloud (mostly LIDAR) data, or both.

2.4.1.3 Scene representation

Scene representation tasks involve the simultaneous mapping of the environment and continuous localization of the AV itself within the environment.

Localization It consists in estimating the position and orientation of the AV with respect to the surrounding environment. These techniques actually belong to the wider set of methods from the area of Simultaneous Localization and Mapping (SLAM), which has been addressing the same range of problems since decades for mobile robots. SLAM algorithms traditionally do not require a priori information about the environment, which allows them to be used anywhere, but the challenging environment in which vehicles evolve, especially in urbanized areas, makes the use of maps [118] a crucial element to achieve a high level of accuracy. Localization is achieved by matching maps with sensory information, such as GNSS or cameras and LIDARs outputs. The fundamental technique being used in this context is visual odometry. Various proposals have also been published to tackle the 3D pose estimation [85].

Occupancy Maps or Occupancy Grids An occupancy grid is a type of probabilistic mask that returns for each cell of a gridded map of the environment the probability that the cell is occupied. This is another popular technique from robotics [119] that is used for localization and mapping in autonomous driving. It can be inferred from the camera and LIDAR data. Techniques for calculating the occupancy grid vary, and have been mainly based on ad-hoc methods, even if currently DL based approaches have also been used, e.g. for subsequent classification of likely objects or the road type [120], [121].

2.4.2. Planning

Planning tasks comprise all the calculations needed to perform vehicle actions autonomously, from route planning to the implementation of an immediate motion trajectory in a given driving situation. They are confronted with the difficulty to evaluate correctly the system predictions: Contrary to perception tasks, where the ground-truth information is usually known and can be compared with predictions, assessing the performances of plann ing systems requires a real world testing in controlled environments, or an evaluation stage in a simulator. Even under these challenging settings, AVs are able to handle most situations, but may fail to take the right decision in scenarios that have not been considered in the data, the model, or the simulations. The reasoning functionalities mainly rely on advanced AI methods for autonomous agents and robotics [122], [123].

2.4.2.1 Route planning

Route planning (or routing), also called global planning, consists in finding the best route between the current position of the vehicle and the destination that is requested by the user. It relies on GNSS coordinates and offline maps that are embedded in the vehicle. The road network is classically represented as a directed graph: nodes of the graph are way points, usually referring to intersections between roads, while edges correspond to the road segments, and are weighted to reflect the cost of traversing along this road, the cost being computed through a metric considering the distance of the segment and/or the time of travel. The problem is then to find the shortest path between two nodes of the graph. The output of route planning is then a sequence of way points that are used to generate the trajectory of the vehicle in the environment.

Several approaches have been developed to address this problem. Routing algorithms are usually relying on specialized heuristics [124] based on graph theory algorithms that have been developed to take into account the size of such graphs (usually several millions of edges) and circumvent the intractability of standard shortest-path algorithms. The efficiency of algorithmic graph solutions makes the use of AI techniques less relevant, albeit ML could be leveraged to adapt in real-time the topology of the graph with external information [125] or provide personalized routes that includes for example touristic sites [126].

2.4.2.2 Behavioural planning

Behaviour planning implies to select what is the most appropriate driving behaviour to adopt for the vehicle, based on the representation of the environment and on the route to follow. Such a behaviour can be described as a sequence of high-level actions. As an example, if the route imposes to turn left at the following intersection, an appropriate behaviour could consist in a sequences of actions such as "Stop the vehicle before the intersection", "Observe the behaviour of vehicles that are coming on the opposite lane and in the crossing lane", "Detect any potential pedestrian that are about to cross the road", and finally "Wait till the path is clear, and then turn left". This decision-making process can be modelled by a finite state machine, where states are the different behaviours of the vehicle, and the transitions between states are governed by the perceived driving context.

One of the key tasks for behavioural modelling is the detection of the driving style of other agents. Driving style designates the various behaviours drivers can adopt while driving, classified with qualifying terms such as aggressive, sporty, calm, moderate, low-skill, or overcautious, to name but a few [127]. Recognising the behaviour adopted by a human-driven vehicle is crucial to understand its dynamics, and is in this respect closely related to the task of tracking the other moving agents. Furthermore, planning systems have to adopt themselves a driving style, possibly giving to the human user a choice between different presets, and find the right trade-off between a conservative driving that could lead to longer journey and aggressive driving that could be unsafe and/or uncomfortable for the passengers. The learning of driving style is an important yet unexplored topic of research, mostly taking advantage of unsupervised approaches [128]–[130] to circumvent the absence of labelled datasets. The use DL has considerably extended the range of modelling capabilities [131], using the vast amount of driving activities recorded by companies to provide simulating environments in which planning models learn to react to different driving scenarios, either by imitating human drivers [132], [133], or through deep reinforcement learning to perform safe and efficient driving [134], [135].

2.4.2.3 Motion planning

Motion planning (or local planning) is responsible of finding the best trajectory of the vehicle in its perceived environment in accordance to the route that has been calculated and the behaviour that has been selected. This consists in the translation of high-level actions into a sequence of way points referring in the coordinates of the perceived environment. This trajectory has to take into account several constraints, such as being feasible by the vehicle (taking into account for example the current speed), being safe, lawful and respectful to other participants present in the environment, as well as ensuring a smooth driving for the passenger.

Traditional approaches have been developed in the robotics community. Typically, the environment is divided into a dynamic grid, where each cell has temporal attributes that are informed by the perception module. The objective is then to find a trajectory between two given cells under multiple constraints, relying on techniques based on graph search, sampling, or curve interpolation [117]. Recently, ML techniques have been employed for local planning, with promising results, in particular in their capacity to avoid erratic trajectories and achieve human-like motions. Generally, these approaches are addressing perception and planning at the same time, either through segmented image data including path proposals [136], or by extracting features from LIDAR point clouds [137]. DL has also been used solely for planning, using RNNs to model sequences of waypoints of trajectories based on a dataset of human motions [138], or reinforcement learning in simulated environments to learn a driving policy that can be extended to real-life situations [139].

ML has achieved tremendous progress in local planning, but has not yet reached a level of maturity sufficient to be implemented in commercial cars. A major limitation is indeed the difficulty to make sure that safety measures are properly learnt, as they cannot be hard-coded in the planning systems as for traditional systems. Nonetheless, their flexibility and they capacity to adapt to unknown situations, provided the context is similar to the one in which the model has been trained, are a strong argument in favour of future deployment of ML based planners.

2.4.3 Control

The control system is responsible for the execution of the trajectory that has been calculated by the planning system by applying commands for the various actuators of the vehicle at the hardware level. Broadly speaking, a vehicle has two types of motions: lateral, controlled by the steering of the vehicle, and longitudinal, controlled by the gas and brake pedals.

Control techniques regroup a set of methods to monitor the dynamics of a system, in order to achieve a given action, while satisfying a set of constraints. Such systems act in a closed loop manner, with an

objective value (e.g. a desired speed) prescribed to the controller, which has the ability to actuate on the systems (e.g. through braking and acceleration), while getting feedback through monitoring to ensure an optimal and stable trajectory of the dynamics of the system. Two popular approaches are Proportional-Integral-Derivative (PID) control [140] and Model predictive control (MPC) [141]. The former consists in continuously calculating the difference between the desired and the measured values of the controlled variable by tuning the relative importance between three different terms, describing corrections to apply to get an accurate and smooth trajectory. The latter relies on predictions of the changes of the controlled value, based on a model of the system. Compared to PID controller, it is costlier in terms of complexity, but allows considering situations where the dynamics has a higher variability, or the delays between actions and feedbacks are higher.

In the case of AVs, the main difficulty lies in the high complexity of the relationships between controlled variables (such as speed or steering angle) and actual commands to actuators. Human drivers, with their experience and their understanding of the physical world, are constantly monitoring the movement of the vehicle, and the different indicators, such as the speedometer, to correct and make sure the behaviour of the vehicle is compliant with their intentions. By doing that, they integrate implicitly parameters as complex as the total weight of the vehicle, the friction forces of the tires on the road, or the wind intensity, through their sensory perception and their modelling of the environment. While it is straightforward to formalize an accurate model between the actuators and the actual behaviour of the vehicle at low speed, additional factors linked to the environment strongly increase the complexity of this model at high speed. Nonlinear control or model predictive control have to be used to take into account this complexity. To this end, ML techniques have already shown great potential to improve the predictive power of control models, albeit as of now they are not deployed in commercial vehicles.

2.4.4 Infotainment and vehicle interior monitoring

AI is not confined to driving functions, and has also been proven useful in infotainment systems and vehicle interior monitoring. These features are starting to be increasingly integrated in modern vehicles [142], [143], offering embedded hardware dedicated to voice recognition, or personal assistant controllable via vocal control and facial expressions.

Human machine interface (HMI) It provides passengers the ability to interact with the car, either to give commands to the driving or entertainment systems for instance, or to receive information, such as the current itinerary. DL is used to provide novel communication vectors, such as speech or gestures. Speech recognition embedded in vehicles are able to understand spoken commands that follows the syntax of normal spoken conversation. Gesture recognition systems are able to interpret common hand gestures, so that gesture based controls become applicable to interactive displays. Recommendation systems to anticipate the choice of users can also be included as part of HMI systems.

Vehicle interior monitoring It consists in monitoring the interior of the vehicle through sensors (e.g. cameras, microphones, temperature sensors, etc.) to ensure the general comfort of passengers. This function has been originally designed to monitor drivers' fatigue, through the monitoring of driver behaviour. Therefore, real-time analysis of biometric factors (e.g. heart rate, respiratory rate, eye blinking, etc.) could trigger a warning alarm to alert the driver. For fully AVs, this function could be used to control the level of comfort, by automatically adjusting sounds, lights, or any additional factors based on predictive models of the well-being of human passengers.

2.4.5 Current trends in AI research for autonomous driving

End-to-end approaches: A general trend in ML is the use of an end-to-end, holistic approach to tackle several problems at the same time. The rationale behind this approach is that developing separate modules tend to be inefficient in terms of computational power, but also may lead to poorer performance, as uncertainties appearing in the upstream section of the driving pipeline are propagated and amplified along all modules. Several variants of this approach exist: a popular one is to consider jointly all tasks of perception or planning, or even all tasks from both modules altogether. Nonetheless, the approach relies on a wide diversity of techniques, ranging from the prediction of driving paths from camera images [136], point clouds, GNSS measurements, and or external information [137], to the prediction of steering commands from the same kind of inputs [144]. Behaviours can also be predicted in an end-to-end fashion from raw pixels [145], [146].

Simulation: The cost of data acquisition has led to the development of many simulators, in order to address large quantities of data. These simulation environments, many of them released as open-

source software, also lowered the upfront investment necessary to do research on AVs, and have been the basis for numerous research works, some of them described in this report. Popular simulators include TORCS [147], CARLA [148] and AirSim [149], which take advantage of graphics engine used in video games to offer a realistic representation of the world. These simulators, as well as others [150] also include tools to customize sensors (e.g. cameras or LIDARs [151]), offer typical driving scenarios to play, and provide easy integration of ML tools for quick development. Other initiatives have been launched to promote autonomous driving research towards students and tech enthusiasts, such as DeepTraffic [152].

2.5 Mapping between automotive functionalities, hardware and software components and AI techniques

As a conclusion of this section, the most important key findings are summarized in the form of three tables. First, a correspondence between the high-level functionalities and the intermediate tasks is given. Then these tasks are mapped respectively with the hardware and software components that have been identified, and finally with the AI techniques.

Automotive Functionality	Detection of roads	Detection of lanes	Detection of moving agents	Traffic sign recognition	Markings recognition	Tracking of objects	Sound event recognition	Localization	Occupancy maps	Routing	Behaviour planning	Motion planning	Trajectory execution
Adaptive cruise control	X		X			X						X	X
Automatic Parking	X	X			X							X	X
Automotive navigation								X		X			X
Blind spot / cross traffic / lane change	X	X	X		X				X		X		X
Collision avoidance systems	X	X	X			X			X			X	X
Lane keeping	X	X	X		X								X
Traffic sign recognition			X	X									
Environmental sound detection							X						

Table 2. Correspondence between high-level functions and low-level functions

Automotive Functionality	Software Components			Hardware Components						
	Perception	Planning	Control	Camera	LiDAR	GNSS	IMU	Radar	Acoustic sensor	Ultrasound sensors
Detection of roads	X			X	X	X				
Detection of lanes	X			X						
Detection of agents	X			X	X			X		X
Traffic sign recognition	X			X				X		
Markings recognition	X			X						
Tracking of objects	X			X	X			X		
Localization	X			X	X	X	X	X		
Occupancy maps	X			X	X			X		
Routing		X				X	X			
Behaviour modelling		X		X	X			X		X
Motion planning		X		X	X	X	X	X	X	X
Trajectory execution			X	X	X	X	X			
Sound event recognition	X								X	

Table 3. Correspondence between low-level functions and hardware and software components

Automotive Functionality	Computer Vision			Sequential Machine Learning			Automated Planning			Control	End-to-End Approaches
	Object Detection	Semantic / instance Segmentation	Vehicle localization	Recurrent models	Markov Models	Filtering models	Classical Planning	Imitation Learning	Policy learning		
Detection of roads		X		X							X
Detection of lanes	X	X		X							X
Detection of agents	X						X				X
Traffic sign recognition	X	X									X
Markings recognition	X	X		X							X
Tracking of objects	X		X	X	X	X					X
Sound event recognition				X							
Localization	X	X	X								X
Occupancy maps	X	X	X	X	X	X					X
Routing							X				
Behaviour modelling							X	X	X		
Motion planning							X				X
Trajectory execution										X	X

Table 4. Automotive functionalities and related AI techniques

3. CYBERSECURITY OF AI TECHNIQUES IN AUTONOMOUS DRIVING CONTEXTS

3.1 Vulnerabilities of AI for autonomous driving

The development of increasingly autonomous and connected vehicles inevitably requires a higher level of computational functionality and connectivity, which, in turn, widen the attack surface and the likelihood of physical and cyber-attacks. Cybersecurity risks in autonomous driving vehicles can have a direct impact for the safety of passengers, pedestrians, other vehicles and related infrastructures. It is therefore essential to investigate potential vulnerabilities introduced by the usage of AI. This section focuses on the general vulnerabilities and security challenges in autonomous driving posed by AI, with a particular focus on ML. It also includes an analysis for specific AI-related vulnerabilities in autonomous cars.

Following consolidated threat modelling practice [153], threats related to AI can be divided into two groups: intentional and unintentional. Intentional threats include those coming from a malevolent exploitation of the limitations and vulnerabilities present in AI and ML methods to cause intended offence and harm. Intentional misuse of AI leads to change of the current cybersecurity landscape by introducing a new class of vulnerabilities and raising the ceiling of potential impacts. The growing use of AI to automate decision-making in a diversity of sectors exposes digital systems to cyberattacks that can take advantage of the flaws and vulnerabilities of AI and ML methods. Since AI systems tend to be involved in high-stake decisions, successful cyberattacks against them can have serious impacts. AI can also act as an enabler for cybercriminals: Cybercriminals can use AI to automate aspects of their attacks, enabling them to launch attacks more quickly, at a greater scale and a lower cost and with higher precision.

Unintentional threats come as side effects of benevolent usages, due to open issues inherent in the trustworthiness, robustness, limitations and safety of current AI and ML methods. Unintentional threats comprise unpredictable malfunctioning, failures or negative aftermaths caused by shortcomings, poor design and/or inner peculiarities of AI and ML. Experimental research and real-settings operations have demonstrated that these methods may suffer from several issues. This includes unfairness of the decision made due to the propagation of biases from data to models and outcomes, opacity of the decision process due to complex model structures and mathematical operations that escape from an ease straightforward interpretation, unsafety due to critical scenarios badly represented or outside the training data fed to the model during the development phases, or challenging reproducibility and verification that can convey a mismatch among real and expected results of ML methods and cause issues when reproducing and investigating the decisional process. These issues affect also the reliability of the methods when used in practice.

The present report focuses on the exploitation of AI vulnerabilities to compromise the integrity and availability of AVs, which belongs to the category of intentional threats. In particular, adversarial ML is discussed, as a prominent field of research linked to cybersecurity of AI, and as an immediate threat for AVs. It is nonetheless worth mentioning the same technical challenges underpin both intentional and unintentional challenges, and improving all aspects described above benefits security and safety of AI systems as a whole. Research in ML is gathering a lot of interest and aggregating substantial community effort, providing advances to increase the robustness and reliability of AI methods in both normal and adversarial settings. A larger overview of AI cybersecurity is discussed in the ENISA AI Threat Landscape [1], its relevance in the larger context of digital transformation is outlined in a JRC report on cybersecurity [235].

3.1.1 Adversarial machine learning

Adversarial ML emerged in 2004 dealing with the robustness of antispam filters [154] and has since then evolved investigating how to challenge and guarantee the security of ML methods and systems [155]. Since then, a large amount of work has been done, suggesting that ML-based systems could introduce further vulnerabilities easily exploitable by skilled attackers. Paradigmatic cases of attacks against ML systems used for pattern recognition in cybersecurity are: submitting a fake biometric trait to a biometric

authentication system (spoofing attack) [156], [157]; modifying network packets belonging to intrusive traffic to evade intrusion detection systems [158], [159]; manipulating the content of spam emails to make them escape spam filters (e.g. by misspelling common spam words to avoid their detection) [160], [161]; manipulating of malware to evade ML-based malware detection; deceiving face recognition systems [162], [163]; taking control of a voice interface system, by way of hidden voice commands, unintelligible to a human listener [164] or the deceit of reading comprehension systems [165]. The possibility to subvert otherwise-reliable ML systems has received considerable attention since 2014, when it was shown that CNNs for object recognition could be tricked by passing them slightly perturbed images [166], [167]. Much effort has been devoted to the topic since then, establishing the subfield of adversarial ML as the most active area of research focusing on the security and robustness of ML systems to adversarial input, especially those relying on DL.

The most common attacks on AI systems can be distinguished between evasion and poisoning attacks. The first type of attacks aims to manipulate what is fed into the AI system in order to induce a system output that serves the attacker's goal. On the other hand, poisoning attacks corrupt the training, so that the resulting system malfunctions in a way desired by the attacker. A big share concentrates on attacks to supervised learning models, including attacks against regression methods [168], [169], SVM [162], and ensembles of classifiers [170]. Vulnerabilities of unsupervised learning models have also been explored, examining possible attacks against clustering methods [171]–[173]. With the greater integration of DL techniques in many critical applications, this area of research has gained much attraction in the last years [174]–[178]. Recently, reinforcement learning models have been probed with respect to vulnerabilities [179]–[181], as a consequence of their reliance on DL models [182]. Other attacks are continuously devised by the research community, for instance against real-time video classification systems [183], or against RNNs [177].

3.1.2 Adversarial examples in computer vision

Adversarial examples are the result of an evasion attack, and consist in tiny perturbations of the input that cannot be detected by human but are leading to a misclassification with high confidence by ML models. Albeit adversarial examples can be found for any kind of inputs, such attacks have been particularly explored for computer vision models. As shown in Figure 8, adversarial examples are typically created by adding a small amount of carefully calculated noise to a natural image. This kind of attack can fool state-of-the-art, highly performant image-recognition models whilst being often imperceptible to humans.

The research on adversarial examples has gradually become a hotspot in the computer vision community, and researchers have constantly proposed new adversarial attack methods. A commonly referred setting for adversarial crafting considers that the adversary's goal is to define a perturbation that, applied to an input image, makes the model misclassify the resulting perturbed image [186]. The ways of generation of adversarial perturbation depends on the adversary's knowledge of the system. A distinction is commonly made between white-box and black-

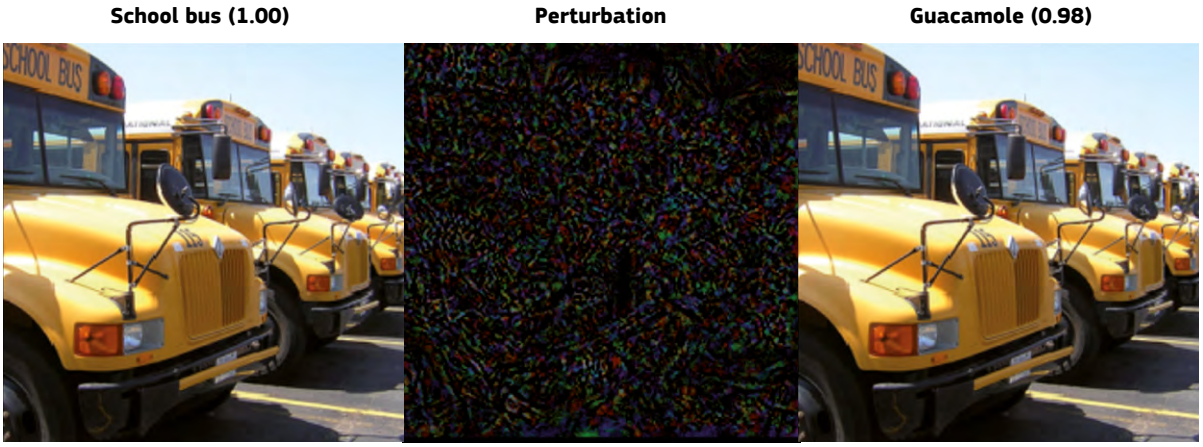

Figure 8. Illustration of an adversarial example using the Basic Iterative Method [184]. The classifier used is Inceptionv3 [71]. The image comes from the validation set of the ImageNet dataset [185]. (Left) Original image, correctly classified as a school bus. (Middle) Perturbation added to the image, with a 10x amplification. (Right) Adversarial example, wrongly classified with high confidence.

box attacks [187]: In white-box attacks, the attacker has a full knowledge of the model that typically includes the parameters of the model and sometimes the data used for the training. Conversely, black-box attacks consider that the attacker has only access to a limited set of pairs of inputs-outputs or is only able to submit its own inputs to the model and gets the corresponding outputs. An intermediate situation, often referred as grey-box setting, considers the same situation as for black-box attacks except the adversary has a limited knowledge about the model, e.g. its training set or the family of models that is employed [188]. The generation of adversarial examples for ML models is most commonly studied from the standpoint of a white-box attack. Black-box attacks have nonetheless been demonstrated in many contexts, and are usually built on top of white-box attacks using substitute models [187], relying on the transferability of adversarial examples from a model to another one, i.e. their capacity to work on a range of close yet different architectures.

The field of adversarial ML has been mainly built on top of computer vision techniques, and in particular of classification models. This report reflects this situation, which is also especially relevant for AVs that are composed of multiple classification systems working on images. It is nonetheless worth mentioning the techniques introduced can be generalized to other types of problems (detection, regression, behaviour modelling, etc.) and data (text, sound, tabular data, etc.), extending the scope of adversarial attacks.

3.1.2.1 Overview of adversarial example attacks

Following the seminal work of Szegedy et al. [167], most techniques proposed to perform adversarial attacks are exploiting the gradients of the model, computed in order to maximise the effect of adversarial perturbations in altering the output of the model. As for the training phase that exploits the same idea to update the model parameters, the minimization of a loss function is used to describe the adversarial problem [189]. The complexity of the problem usually makes the resort to an optimization algorithm necessary to get a satisfying solution, which is mostly based on a classical technique called Projected Gradient Descent (PGD) [190].

A first line of research consists in improving the optimization process at play during the attack to make it practical for large dataset and allows for better solutions. This is done either by finding better terms to include in the loss function (i.e. that are both a good approximation of the objective one want to achieve and have properties that improves the minimization process), or improving the algorithm used to minimize it (i.e. finding the iterative steps that will converge to a good solution). Fast Gradient Sign Method (FGSM) [191] relies on a single optimization step that includes only the signs of the gradients. Basic Iterative Method (BIM) [184] extends this approach by applying FGSM iteratively, increasing the chance of successes of attack by crafting more complicated perturbations [192], [193]. Other approaches, like Jacobian Saliency Map (JSMA) [194], rely on the localization of the salient pixels of the images and focus on these specific areas to craft the perturbations. C&W attacks [195] add multiple refinements to previously mentioned techniques to increase the chance of success, and in particular to bypass several defensive mechanisms that have been suggested to counter adversarial attacks.

A second line of research that is complementary and usually considered simultaneously in the design of the optimization problems concerns the choice of the constraints to apply on the perturbations. Most techniques are typically trying to find the minimal perturbation possible, in order to make it less perceptible by a human auditor. This is often expressed as the average over all pixels of the intensity of the perturbation, the maximal intensity, or the number of pixels that are perturbed. Depending on the type of constraints, the optimization algorithms will be chosen accordingly to achieve a good convergence, i.e. make sure that the whole process returns a perturbation that is indeed a good minimum for the considered loss function. As an example, the DeepFool attack [175] computes a minimal norm adversarial perturbation for a given image in an iterative manner, in order to find the decision boundary closest to the clean input image and find the minimal adversarial sample. Some attacks are also considering very specific setting, like the One Pixel attack [196] that constraints the perturbation to affect only one single pixel. Other approaches exist to address different contexts and applications. The Houdini attack [197] works on non-differentiable loss functions, and has been proven useful in domains such as natural language processing, while the Zeroth Order Optimization (ZOO) attack [198] has been used in black-box settings to estimate gradients based on outputs. A current trend is the use of generative models [199], [200] to synthesize adversarial examples entirely from scratch. While these approaches do not allow modifying existing images, their capacity to generate realistic samples wrongly classified is getting more and more problematic, and is closely related to the growing use of deepfakes [201].

Similar approaches have been followed to generate adversarial examples for other types of data, not

coming from cameras but from other sensors that can be found on autonomous vehicles. Examples include attacks on 3D points clouds outputted by LIDARs [18], [202]–[204], on radars [205] or ultrasonics sensors [206]. The paucity of standard datasets makes this area of research less fruitful compared to traditional RGB images, but the democratization of devices and the development of methods following the same concepts as for traditional imagery (e.g. transferability of adversarial examples between 3D point cloud models [207]) will increase the efficiency of adversarial attacks.

3.1.2.2 Physical adversarial examples

Adversarial attacks described above mainly refer to a setting where attackers have the ability to update the inputs with the only constraints that perturbations were imperceptible for a human supervisor. This setting has some limitations when data are processed after acquisition, as it often happens in computer vision: a model may be trained to recognize a certain type of objects using a large collection of digital images, but the end application would involve a camera acquiring and processing on-the-fly images. In the last couple of years, it has been extensively demonstrated that adversarial examples could be transferred to the physical world [208], [209]. This is mainly done through the alteration or the creation of objects with specific features that are misclassified by a DL model trained to recognise them after acquisition by a camera.

The generation of successful physical adversarial attack is particularly challenging due to the loss of sensitivity of adversarial perturbations when they are subject to minor transformations [210] either happening in the physical environment, such as lighting variations, change of angles, motion blurring, etc. or in the acquisition phase, such as filtering or resizing. An interesting workaround consists in including these transformations in the optimization scheme, either in the iterative process like for the expectation-over-transformation (EOT) attack [211] that alternates between gradient steps and random transformations, or adding terms in the loss function that constrain the perturbations to be physically realisable [20].

3.1.2.3 Countermeasures to adversarial attacks

In parallel of the development of adversarial attacks, defensive measures to make these systems less vulnerable have been developed [212]. This has led to a cat-and-mouse situation between offensive and defensive adversarial ML, which is for now taking place in the ML research community, but is expected to develop into a more typical cybersecurity situation between adversaries and defenders in the future. A short overview of the main techniques developed in the ML community is given in the following. Adversarial attacks may be counteracted by approaches to make the ML model more robust and resilient to adversarial examples. The solutions presented in the literature, especially in the last years in relation to DL models, may be categorised in two strategies: acting on the data or acting on the model.

Acting on data can take place at training of inference time: Adversarial training [213] consists in introducing adversarial examples generated using classical attacks into the training dataset to improve the robustness of the model [154]: the model integrates the variability added by the adversarial perturbations in the model. The resulting model is less subject to adversarial attacks, and is able to correctly handle adversarial inputs generated using substitute models [214]. Once the model is trained, a second action to mitigate the risk of adversarial attacks consists in performing data sanitization, which is used to detect and reject samples that are too far from the training data distribution [215]. For computer vision DL models, the use of generative networks has also been proposed to identify and reject adversarial examples prior submitting them to the model [216], [217]. These two approaches techniques have nonetheless strong limitations, adversarial training protects the model only against a limited set of attacks, and has a significant impact on the training performances, both in terms of accuracy and training time. As for data sanitization, the full check can also be computationally expensive, with limited effectiveness depending on the kind of attacks.

Actions on the model mainly aim at increasing the robustness of the model by making changes in the training algorithm. In this frame, a classic technique called regularization appears to improve the generalization capacity of the model by adding penalty terms to the cost function forcing the parameters of the models to exhibit desirable properties, like smooth decision boundaries, and to increase the resistance of the model to attacks on unknown data [218], [219]. Defensive distillation [220] is also used to smooth the outputs of the model, and avoid hard decision boundaries that are exploited by many adversarial attacks. These defences are however heuristic, with no formal guarantees on convergence or robustness properties. Formal verification approaches [221], [222] are being more and more popular in the research community, and have demonstrated convincing results. Their use in AI systems for AV is nonetheless premature due

to the low scalability of these techniques. Finally, ML ensembles have also been exploited to improve security against evasion attempts, e.g. by implementing rejection-based mechanisms or secure fusion rules [170], [223], [224], here again with an additional computational cost.

3.1.3 AI-based physical attacks against autonomous vehicles

Several examples of physical adversarial attacks on AI components of semi-autonomous cars were reported in recent years. DARTS [225] (Deceiving Autonomous caRs with Toxic Signs) targets the traffic-sign recognition functionalities of autonomous cars. The method includes a pipeline for upscaling adversarial perturbation in such a way it becomes printable, and has been evaluated on real-size printed signs to fool a classifier getting images from a front-facing camera in a real vehicle. The adversarial creation pipeline proposed for DARTS has been extended [226] to deceive a commercial car perception system in real-world driving conditions, with improved random augmentation techniques and the ability to create perturbations that are tailored to speed limit traffic signs and, therefore, less perceptible to a human viewer. The pipeline allows for robust production and evaluation of printing-size adversarial signs in black-box models. Spoofed and clean signs were positioned around the track, and were perceived by the traffic sign recognition system of the car driving around the track. Results showed that the altered signs were not only misclassified, but also caused some unexpected behaviours of the vehicle.

Another example of an attack consisted in deceiving Tesla cars into accelerating well past a speed limit [227]. By slightly elongating using black tape the middle line in the "3" on a 35-mph (around 56 km/h) speed sign, the system predicted a speed limit of 85 mph (around 137 km/h). In another work [228] focusing on the popular external ADAS Mobileye, the researchers injected spoofed traffic signs to assess the influence of environmental changes (e.g. changes in colour, shape, projection speed, diameter and ambient light) on the outcome of an attack. To conduct this experiment in a realistic scenario, they used a drone to carry a portable projector, which projected the spoofed traffic sign. Their experiments show that it is possible to fool Mobileye so that it interprets the drone projected spoofed traffic sign as a real traffic sign.

A research group from Tencent Keen security lab performed a comprehensive study of reverse engineering for finding security flaws in a Tesla car [229]. Among other things, they found weaknesses in the perception systems of the vehicle. They were able to trigger the auto wipers by projecting noise on an electronic display placed in front of the vehicle, thus fooling the visual sensor of the system. They also investigated the lane detection system: it was demonstrated that after application of aggressive blur to a traffic lane the perception system might not detect it, and that fake lanes might be produced by placing certain stickers on the road (the latter was not demonstrated yet in real driving conditions). In this situation, a human driver would have probably noticed the perturbation, but would have relied on common sense to react properly.

A recent work has also demonstrated that the steering angle predicting systems of an autonomous car is vulnerable to adversarial evasion attacks at operation time [230]. The authors have adapted the Carlini & Wagner attack to change the predicted steering direction.

3.2 Attack scenarios related to AI in autonomous driving

Considerable research effort is being invested in identifying AI security issues and vulnerabilities for AVs, recommending potential mitigation techniques, as well as highlighting the potential impacts on the vehicle itself and related infrastructures becoming compromised. Various threats associated with the different sensors, controls, and connection mechanisms have been identified. In addition to the vulnerabilities specific to ML systems discussed in the previous section, AI-related security issues are taking advantage of the more classical hardware and software vulnerabilities present in digital systems, extending standard attack vectors. More precisely, some of these security issues and vulnerabilities usually mentioned include:

- Sensor jamming, spoofing and blinding/saturation: sensors may be blinded or jammed. In this way, the attacker may manipulate the AI model, feed the algorithm with erroneous data or intentionally provide scarce data and thus diminishing the effectiveness of automated Decision-making. Stemming from the first attempts [17], recent works have demonstrated for instance the possibility to saturate [18] or spoof [202] LiDAR sensors and its underlying ML method for data interpretation.
- DoS/DDoS attacks: disrupting the communication channels available to an AV makes it essentially blind to the outside world. It has a direct impact on its availability and hinders operations needed for autonomous driving. The objective of DDoS attacks is to disrupt such communication channels.

- Manipulating vehicle communications: hijacking and manipulating communication channels have a severe effect on autonomous driving operations, allowing an adversary to modify transmitted sensor readings or falsely interpret messages coming from road infrastructure.
- Information disclosure: given the abundance of (personal and sensitive) information stored and utilized by vehicles for the purpose of autonomous driving, including critical data on the AI components , a particular motivation emerges for potential adversaries to gain access to this type of information and cause a data breach.

3.2.1 Attack scenarios

Five hypothetical scenarios are presented in this section, to illustrate the exploitation of AI vulnerabilities in an automotive context using both classical cybersecurity and AI-specific vulnerabilities.

AUTONOMOUS VEHICLE COMPROMISE

DESCRIPTION	
Adversaries introduce physical perturbations on the road markings to deceive the model into perceiving wrong information about the environment. This includes alterations, placement of stickers, or projection of light on the painting of the road lanes or on road signs (stop signs, speed limit signs, etc.). These carefully crafted patterns lead to a misclassification of objects or symbols by the perception component, and subsequently to misbehaviours of the AVs.	
IMPACT	
Medium – High: The impact depends on the target markings, and the role that it plays in other autonomous driving functions. Misclassification of markings can easily generate safety issues, triggering misbehaviours in autonomous navigation functions endangering road users' safety and leading to driver, passenger, or pedestrian deaths.	
EASE OF DETECTION	**CASCADE EFFECT RISK**
Easy – Medium – Hard: Depending on the nature of the attack, the alterations could be detected easily, or on the contrary remain undetected by human eyes before an accident occurs.	**Low:** The perturbation is local, and may affect only the cars passing by the modified marking.
ASSETS AFFECTED	**STAKEHOLDERS INVOLVED**
Markings recognition algorithms Sensors Vehicle functions	OEMs Road infrastructure
ATTACK STEPS (SAMPLE BASED ON A REAL-CASE ATTACK SCENARIO)	
1. The attacker first analyses the capabilities of the targeted versions of cameras and AI-based image classifier and designs an adversarial attack able to alter the outputs. This phase may require trying multiple perturbation patterns or display parameters. The attacker needs to perform some physical experimentation as well to ensure that the attack will succeed. 2. At a next step, the attacker performs the alteration of the targeted marking or traffic sign to cause misclassification by the AV. 3. Due to the added perturbation, targeted autonomous cars passing by the altered marking or traffic sign will erroneously classify it into the attacker's chosen class (e.g. interpret a stop sign as a speed limit sign) and react accordingly (e.g. reduce speed instead of stopping the vehicle).	
RECOVERY TIME / EFFORT	**GAPS AND CHALLENGES**
Medium: Sensor fooling attacks can go unnoticed. Once detected, modified markings or traffic signs can be reverted in hours.	Markings and traffic sign authentication Design of robust AI models Collaboration of vehicles
COUNTERMEASURES	
1. Hardening against adversarial examples. 2. Use of hardware redundancy mechanisms. 3. Use of data redundancy mechanisms, such as multiple sensors. 4. Perform data validation, for instance by comparing sign information collected by sensor with information from digital maps stored in the vehicle. 5. Use V2X communication to receive read sign information.	

Attack scenario 1: Adversarial perturbation against image processing models for street sign recognition and lane detection

AUTONOMOUS VEHICLE COMPROMISE

DESCRIPTION

An adversary discovers a remotely exploitable vulnerability in the vehicle's head unit (HU). The attacker exploits this vulnerability over Internet to compromise remotely the HU of vulnerable vehicles. Once inside the HU, the attacker performs lateral movements gaining access to the in-vehicle network. On the other hand, the attacker may have direct access to the internal network during the car maintenance. If the car's internal network does not authenticate well its components, injecting a tampered module can do the trick better than hijacking an internet connection.
From that advantageous point, the adversary performs a man-in-the-middle attack on the state representation of the environment outputted by the perception module. We assume that the attacker can only add small perturbations to the state values to avoid detection. To select the right perturbations, an adversarial attack on the reinforcement learning model used to select the right behaviour to adopt considering the state of the environment is designed, leading to a change of behaviour of the autonomous cars.
Examples of attacks include replacing a braking command emitted when a stop sign is detected, by an acceleration command and allowing for a turn even if an obstacle is present on the trajectory.

IMPACT

High: The impact depends on the specific misbehaviour generated in the system. If the systems in charge of the vehicle actuators are targeted, the potential impact is very high, as the vehicle might be driven to perform unsafe manoeuvres (like emergency braking).

EASE OF DETECTION	CASCADE EFFECT RISK
Difficult: the adversarial ML attack is carried out within the in-vehicle network, where the attacker has a more fine-grained control over the AI inputs and their actions go easier un-noticed by the human operator.	**High:** In this scenario, the initial entry point of the attack is a remotely exploitable vulnerability that can be triggered from Internet. The adversary could easily automate this, potentially affecting an entire fleet of vulnerable vehicle at international level.
ASSETS AFFECTED	**STAKEHOLDERS INVOLVED**
Motion planning algorithms Vehicle functions	OEMs Road infrastructure Mobile operators

ATTACK STEPS (SAMPLE BASED ON A REAL-CASE ATTACK SCENARIO)

1. The attacker first identifies and finds the way to exploit remotely a vulnerability on a HU service reachable from Internet.
2. Once the HU is compromised, the attacker finds the way to move laterally and gain access to the in-vehicle network.
3. Man-in-the-middle attacks are used to hijack the data input by AI components.
4. The attacker analyses the AI model used and launches an adversarial machine learning attack manipulating the planning module output.

RECOVERY TIME / EFFORT	GAPS AND CHALLENGES
High: If no Over-The-Air (OTA) update mechanism is in place, the patching of vulnerable vehicles can take a considerable amount of time and effort.	Agile patching mechanisms Robust ML

COUNTERMEASURES

1. Follow well-known cybersecurity principles, to protect against the elements of the cyber chain that do not relate to AI.
2. Hardening against Adversarial Machine Learning.
3. Use of hardware redundancy mechanisms.

Attack scenario 2. Man-in-the-middle attack on the planning module.

AUTONOMOUS VEHICLE COMPROMISE

DESCRIPTION

Autonomous cars in circulation are indeed constantly sending information to the company, in particular in edge-case situations where the model encountered a high uncertainty on the decision to take. Edge cases can be exploited by malicious actors to inject unexpected behaviours inside models using update of the AI models of AVs regularly done by manufacturers. To do that, an attacker could create a sign that has the shape of a stop sign, and has the word "SHOP" written on it. Humans would not consider the sign as a real traffic sign, but AVs may consider it, and adopt a safe approach and stop, while triggering an anomaly. The anomaly could be corrected by a human operator, associating in the model the traffic sign "SHOP" with the action "DO NOT STOP". Repetitions of the same operation with different vehicles by the attacker could be done to increase the likelihood of integration in the model. After deployment of the update on vehicles, the attacker could simply put a sticker to replace the "T" into a "H" on any stop sign to cause accidents.

IMPACT

High: The impact depends on the specific misbehaviour generated in the system. If the systems in charge of the vehicle actuators are targeted, the potential impact is very high, as the vehicle might be driven to perform unsafe manoeuvres (like emergency braking).

EASE OF DETECTION	CASCADE EFFECT RISK
Easy - Difficult: the poisoning attack could be easily detected with a robust validation process of anomalies returned by autonomous cars. Once validated, the detection inside the model could be difficult.	**Medium:** Once the update deployed, the entire fleet of vulnerable vehicle at international level would be affected.
ASSETS AFFECTED	**STAKEHOLDERS INVOLVED**
Decision Making algorithms Vehicle functions	OEMs Road infrastructure

ATTACK STEPS (SAMPLE BASED ON A REAL-CASE ATTACK SCENARIO)

1. The attacker first identifies a pattern that is close enough to a sign to be misinterpreted by the system while being at the same time easily identifiable by a human operator.
2. The attacker undertakes driving experiences including the pattern, in such a way that it is associated to an action that differs from the one of the traffic sign.
3. The attacker relies on the fact that the system will raise a warning that will be operated by a human operator that will consider the pattern as a false positive and update the model to take into account this edge case.
4. After update of vehicles, the attacker modifies the targeted traffic sign to deceive the AV into doing a wrong action.

RECOVERY TIME / EFFORT	GAPS AND CHALLENGES
High: Detecting the erroneous update could take some time.	Agile patching mechanisms Robust machine learning

COUNTERMEASURES

1. Hardening against Adversarial Machine Learning.
2. Use of hardware redundancy mechanisms.
3. Authentication of the signs to the vehicle

Attack scenario 3. Data poisoning attack on stop sign detection.

AUTONOMOUS VEHICLE COMPROMISE

DESCRIPTION

In this scenario, an adversary may discover a remotely exploitable vulnerability and deploy malicious AI firmware from back-end servers. This could be initiated by OEMs employees (e.g. developers) or by external attackers capable of penetrating back-end servers. Malicious OTA (Over-the-air) updates of the AI models could then be executed so that AVs think it is a legitimate one, as it is initiated from a trusted server. The attack might be used to make the AI "blind" for pedestrians, by manipulating for instance the image recognition component in order to misclassify pedestrians. This could lead to havoc on the streets, as autonomous cars may hit pedestrians on the road or crosswalks. Given that such OTA updates are being pushed at scale to the entire fleet of vehicles of particular model/brand, it is easy to envisage that the scenario involving the entire fleet may have detrimental safety impact.

IMPACT

High – Crucial: Remote servers might communicate with numerous vehicles at the same time. Thus, compromising the AI models of such a centralised server could affect the entire ecosystem, including passengers' safety.

EASE OF DETECTION	CASCADE EFFECT RISK
Medium: Remote servers should have enough resources to implement advanced monitoring techniques. However, the deployment of many remote servers increases the attack surface to be protected.	**High:** Such attacks are highly-scalable as they can be executed remotely and based on the compromised AI models they can affect a fleet of vehicles instantly.
ASSETS AFFECTED	**STAKEHOLDERS INVOLVED**
OEM Back-end system Software and Licenses OTA Updates Vehicle functions Information (User, Device, Keys and Certificates)	OEMs

ATTACK STEPS (SAMPLE BASED ON A REAL-CASE ATTACK SCENARIO)

1. To perform this attack scenario, the attacker needs first to penetrate the targeted OEM back-end server. This may be carried out by leveraging a known vulnerability of used software, a misconfiguration on the server side or by spoofing the administrator account for instance.
2. Once the attacker gets access to the OEM back-end server, the attacker can request the execution of an OTA firmware update of the AI models or the image recognition component for a fleet of given vehicle models/brand. To this end, the attacker follows the same steps required to perform a legitimate OTA firmware update.
3. Upon receiving the OTA update request, vehicles acknowledge and accept the request as it is initiated by a legitimate OEM server.
4. Next, the attacker uploads a rogue firmware of the AI models on the OEM back-end server and launches the OTA update process to deploy this firmware.
5. Once the rogue firmware is installed on smart cars, the attacker can take remote control of a fleet of vehicles by exploiting a backdoor introduced in the rogue firmware or by adversely affecting the expected behaviour of all vehicles.

RECOVERY TIME / EFFORT	GAPS AND CHALLENGES
Medium – High: Depending on the nature of the deployed firmware, cancelling the update by returning back to the retro version can be challenging if the attacker was able to change AI models update related information (e.g. certificates, policies) or the image recognition component that utilizes real-time data. Use of logging can help to identify the attack origin.	Lack of validation mechanisms for the inputs of the AI system Lack of awareness and knowledge Lack of a secure boot process Lack of proper product lifecycle management

COUNTERMEASURES

1. Regularly assess the security controls and patch vulnerabilities.
2. Deploy Intrusion Detection Systems (IDS) at vehicle and back-end levels.
3. Introduce a new device or software change into the vehicle only according to an established, accepted and communicated change management process.
4. Consider establishing a CSIRT.
5. Apply security controls at back-end servers.
6. Establish an incident handling process.
7. Incident report to back-end servers.
8. Conduct periodic reviews, of authorization and access control privileges for instance.
9. Software authenticity and integrity checked before installation.
10. Use of secure OTA firmware updates.
11. Protect OTA update process.
12. Use of secure boot mechanisms.
13. Application of security controls to back-end servers.
14. Apply least privileges principle and use individual accounts to access devices and systems.
15. Maintain properly protected audit logs.
16. Allow and encourage the use of strong authentication mechanisms.

Attack scenario 4. Attack related to large-scale deployment of a rogue firmware after hacking OEM back-end servers

AUTONOMOUS VEHICLE COMPROMISE

DESCRIPTION

An adversary may jam wireless sensor and communications producing radio interferences to disrupt wireless networks so the sensors cannot receive messages and in general, vehicles cannot emit or receive V2X messages. Additionally, an adversary may also spoof the communications by emitting false signals (e.g. GNSS-like signals, with the intent to produce false location-based information in the victim receiver). The malicious signals may also be exploited to affect adversely the communication channels of wireless sensors. For example, the objective in the latter case may be to deplete battery life or even to jam the communication channel so that the sensors will not be able to send back their readings. Both examples have a direct impact on availability. This will cause problems to the AI models that depend on the targeted sensor and hence in the related functionalities of the vehicle. On the other hand, in case of GNSS spoofing, the AI algorithms are fed with purposefully erroneous data and false decisions will be taken regarding the vehicle functionalities.

IMPACT

High – Crucial: Modern vehicles are fully equipped with a multitude of sensors in order to be able to perform autonomously all driving functions (e.g. sensing, detecting objects, etc.). Thus, through sensor jamming an adversary may inject unwanted signals into the communication channel and block/disrupt the connection of sensors with the related AI algorithms. In the case of GNSS spoofing, the AI models are fed with false and potentially malicious data that affect the decision-making processes and the relevant functionalities of the vehicle including passengers' safety.
With this type of attack, either jamming and/or spoofing, the attacker may influence the control of the AV. This may lead, for example, to different situational awareness understanding, provoke false collision warnings, choose wrong location/positioning of the vehicle and thus generate safety issues.

EASE OF DETECTION

Medium: A jamming and/or spoofing attack does not require any advanced type of hardware or software to be mounted. Typically jamming and/or spoofing attacks are hard to detect, they do however have detrimental effects on the functionality of the vehicle.

CASCADE EFFECT RISK

High: Sensor networks are prone to jamming mechanisms. These signals can entirely engage the channel so that authentic communications cannot take place or the packets in transmission be corrupted. This attack affects the AI models, the decision algorithms and hence the vehicle functionalities. Spoofing can impair the AI models due to the injection of unreliable data.
Such attacks are highly-scalable as they can be executed remotely and affect a variety of vehicles functionalities.
In the context of this type of attacks, other vehicles that communicate with the attacked vehicle may be also affected since the V2X communications are influenced and in case of spoofing wrong messages may lead to unwanted collisions.

ASSETS AFFECTED

Software (e.g. AI models, decision making algorithms)
Vehicle functions
Sensors for AVs
Communication systems
GNSS
Mobile networks/systems

STAKEHOLDERS INVOLVED

OEMs
Service providers

ATTACK STEPS (SAMPLE BASED ON A REAL-CASE ATTACK SCENARIO)

1. The adversary identifies security vulnerabilities in sensors and GNSS signals.
2. The adversary exploits these vulnerabilities remotely by injecting unwanted signals into the communication channel or disable sending/receiving messages. Moreover, spoofers overpower relatively weak GNSS signals with radio signals carrying false positioning information.
3. Once the sensor is compromised, the attacker can block the sensors (data are blocked or disrupted from successful transmission) and hence affect the functionality of the decision algorithms that the vehicle uses to perform the corresponding functionalities (e.g. obstacle detection, lane departure etc). Additionally, once the GNSS signal has been spoofed and the vehicle starts receiving erroneous data, then the AI techniques that are based on positioning functionalities are adversely affected.
4. The attacker may take over control of the AV. For example, this may lead to different situational awareness understanding, provoke false collision warnings, choose wrong location/positioning of the vehicle and thus generate safety issues.

RECOVERY TIME / EFFORT

Medium – High: Depending on the nature of the attack recovery time depends on the vehicle and service providers' ability to identify, isolate and address the attack. The response time for the resolution of this attack will be proportional to the time taken to resolve the situation. Having the required technical tools to identify fake signals may significantly reduce recovery time and efforts.

GAPS AND CHALLENGES

Lack of validation mechanisms for the inputs of the AI system
Lack of awareness and knowledge
Lack of authenticity
Lack of encryption
Lack of security by design
Lack of information sharing

COUNTERMEASURES

1. Power measurements (e.g. higher transmitted power).
2. Advanced interference mitigation technologies.
3. Building security directly into the GNSS satellites (authentication services).
4. Receiver featuring spoofing/jamming detection capabilities.
5. Security by design in sensors and receivers.
6. Regularly assess the security controls and patch vulnerabilities.
7. Strong user authentication mechanisms.
8. Deploy Intrusion Detection Systems (IDS) at vehicle.
9. Disaster recovery plan.
10. Consider establishing a CSIRT.
11. Establish an incident handling process.
12. Apply least privileges principle and use individual accounts to access the vehicle systems.
13. Maintain properly protected audit logs.
14. Analyse if possible attack modes and models in order to develop defence techniques.

Attack scenario 5. Attack related to sensor/communication jamming and GNSS spoofing.

3.2.2 Illustration: Fooling a traffic sign recognition system

A short experiment is provided to illustrate the implementation of an attack performed in the context described in attack scenario 1. We implemented two practical adversarial attacks of an AI model for traffic sign recognition (TSR), responsible for the detection and the identification of traffic signs along the road.

In this illustration, we develop a custom detection system based on the DL based architecture YOLO [76] to perform traffic sign recognition, using an implementation of the YOLOv5 framework [231]. For our application, the model is trained on the German Traffic Sign Detection Database (GTSDB) [106]. The model is tested to make sure the performances of the model were good on previously unseen images. We implemented two attacks to fool the outputs of the TSR system. As described in Section 3.1, a distinction is made between adversarial attacks operating at the level of digital systems (referred in the rest of the section as "digital context") and those applied in the physical environment (referred as "physical context").

3.2.2.1 Overflow attack in the classical context

The overflow attack consists in making the system outputting a large number of detections of signs in the current frame. It is performed in the digital context, where adversaries have access to the system in which the AI component is evolving, and aim to update the numerical values that are being inputted to the detection systems. This happens if the attacker gained access to the internal system, either remotely (e.g. using network access) or physically (e.g. using a vulnerability in the infotainment system), and installed a malicious piece of software running on the internal computer of the car. Here, we consider that the attacker is looking to apply the minimal perturbation possible to make the adversarial image looks similar for a human user.

Figure 9 shows the result of the attack on a driving scene extracted from the GTSDB dataset: more than 100 signs are detected all over the image with high confidence, in contrast with the two signs present in the scene. This attack impairs the availability of the TSR system that can either affect the responsiveness of the autonomous system if not handled correctly, or make the vehicle ignore actual signs present on the road, leading to wrong behaviours of the vehicle.

3.2.2.2 Class spoofing attack in the physical context

Class spoofing consists in making the system outputting a different category for the signs that are detected, the localization of the sign remaining identical. In the physical context, the attacker can only alter the environment in which the car is evolving, for example adding stickers, projecting light, or physical altering signs. The goal is to cause the TSR system to produce invalid, yet plausible output, and thus to induce a wrong behaviour of the vehicle. The system cannot detect the attack, reducing its integrity. The car may adopt a behaviour prescribed by the attacker. In the physical context, the attack can only alter the environment in which the vehicle evolves. For practical reasons, this is translated in this experiment as a constraint that the attack can only modify a few pixels of the image on a sign, the same way a sticker would do.

Figure 9. Visualization of the overflow on a TSR model. (Top) Normal output of the TSR system: the two signs are detected as expected. (Middle) Amplified intensity of the perturbation added to each pixel of the image. (Bottom) Output of the TSR system on the adversarial image: more than 100 signs are detected with high confidence.

Figure 10. Visualization of the class spoofing attack on a TSR model. (Top) Normal behaviour of the model: the "keep right" sign is correctly detected. (Bottom) Output of the TSR system on the adversarial image: the sticker-like perturbation on the sign makes the model incorrectly classify the sign as "priority road".

Figure 10 shows the result of the attack on a driving scene extracted from the GTSDB dataset: adding a sticker-like perturbation located on the sign fools the TSR system that correctly predicts the localization of the sign but misclassifies the sign. This attack impairs the integrity of the system, with potential high impacts on the behaviour of the vehicle in the driving environment.

4. AI CYBERSECURITY CHALLENGES AND RECOMMENDATIONS FOR AUTONOMOUS DRIVING

4.1 Systematic security validation of AI models and data

Data plays an important role when building and validating AI systems, at the core of the learning process of ML models. AVs have multiple sensors collecting each second millions of values describing the environment according to various modalities. These large sets of data are feeding complex AI models that are dynamic in nature. In this context, systematic data validation is of paramount importance in order to prevent unexpected behaviour due the wider variety of situations that vehicles may encounter in the real world, including attacks based on the alteration of inputs such as poisoning and evasion attacks. Companies and research groups are not only relying on real-world static data sets, but also make use of simulation environments to get large volumes of realistic yet simplified data, adding another layer of concern in terms of security. More globally, the definition of data governance adapted to the particularity of data used in autonomous driving should be implemented to understand, among others, who owns the data, who has access, or the appropriate usage of the data.

A particularity of AI models is that they can change their behaviour overtime, implying that security and robustness assessments do not just take place at a given point in time during their development, but instead should be systematically performed throughout the AI model lifecycle. This is of particular importance when considering the proliferation and use of pre-trained models from third parties, and the fact that AI models are constantly learning from newly acquired sets of data. The systematic validation of AI models is a challenging issue for the cybersecurity of AVs, ensuring security in the AI systems in autonomous cars to make sure model updates do not add vulnerabilities that may be exploited by attackers.

In this context, it is important to ensure that the security and robustness of model updates is systematically assessed and tested, as part of a broader systematic validation and testing process to ensure their quality and reliability in relation to data dependencies, model complexity, reproducibility, testing, and changes in the external world. This also includes the data used by the ML models, which may eventually contain unexpected patterns, not presented in the training datasets, unintentional (change of environments, etc.) or be intentionally altered to conduct a cyberattack.

RECOMMENDATIONS

- Establish monitoring and maintenance processes for the AI models either proactive or reactive.

  The proactive monitoring works to identify how to improve continuous learning to provide software updates. The reactive approach entails detecting a wrong output and identifying its causes to understand how the method or the outputs can be rectified. In this context, the same situation can be checked before or after software updates to check if there is a faster way to take decisions.

- Conduct systematic risk assessments considering specifically the AI components throughout their lifecycle.

- Adopt resilience mechanisms preparing alternative plans and incident response activities in case of incidents.

- Establish feedback loops of testing vehicle operations as a continuous monitoring process and lessons learned activities.

- Establish audit processes to support forensic analysis after incidents and address relevant concerns for the future.

  For example, keep audit trails to later check how a decision was made and perform post incident analysis, keep logs of serving data since data might delay things and lead to accidents. This is related both to information security incidents and traffics accidents.

- Introduce additional validation checkpoints to limit the impact of erroneous data.

For example, potential solutions of additional validation mechanisms for continuous validation of data.

4.2 Supply chain challenges related to AI cybersecurity

The security of the software and hardware supply chain is of paramount importance in cybersecurity. The supply chain should be strong, capturing all the involved parts in order to ensure security of the software. Supply chain management is a well-known challenge acknowledged by the majority of involved actors and stakeholders. The absence of proper security policies and sufficient strategies across the supply chain of AI components results in a lack of resilience and the presence of potential security breaches in systems. Ensuring proper governance of security policy across the supply chain requires involving stakeholders as diverse as developers, manufacturers, providers, vendors, aftermarket support operators, end users, or third-party providers of online services.

Recently, the situation has become even more complicated as AI systems are becoming increasingly involved in autonomous vehicles, and have an additional impact on the supply chain and its complexity. Security aspects in all the phases of the AI lifecycle introduce new security risks in the automotive supply chain. For example, checking for security issues (intentional such as backdoors and non-intentional) in pre-trained models is challenging considering the complexity and the opaqueness of AI models. Besides, the distinct open-source culture in ML limits the tracing of such assets, pre-trained models being available online and widely used in ML systems, without guarantee on their origin.

Another particularity of the supply chain security issue in AI for autonomous driving is connected to the specific way in which the automotive industry works with respect to the digital components of the vehicle. Whilst there are new players (e.g. Tesla) that integrate themselves the electronic control units (ECUs), most manufacturers rely on ECUs from third parties, resulting in a vehicle having dozens of ECUs from several manufacturers [232].

Accordingly, security processes of the supply chain should capture the specific AI features and become dynamic and flexible in every potential change. Depending on the architecture of autonomous cars, highly accessible components and lack of robust AI models may entail significant concerns with respect to cybersecurity. It could be very beneficial to use secure embedded components to perform the most critical AI functions, similarly to the usage of hardware security module for cryptography. Having unauthorized access to an unsecured element could lead to threats for the whole automotive ecosystem.

Cybersecurity is a shared responsibility between all stakeholders, including OEM, Tier 1 and Tier 2 entities, who should address security concerns to mitigate the different risks and ensure people's safety. The current draft of the UNECE regulation [2] specifies that OEMs, suppliers and service providers should consider cybersecurity concerns and implement appropriate security controls. However, the security regulations and good practices should take into account the specific features and the relevant impact of the involved AI models.

RECOMMENDATIONS

- Establish a proper AI security policy across the supply chain, including third-party providers.
- Ensure governance of AI security policy across the supply chain.
- Identify and monitor potential risks and threats related to AI in autonomous driving.
- Develop an AI security culture across the supply chain, involving all the stakeholders.
- Request compliance with regulations in the automotive sector across the supply chain.

4.3 End-to-end holistic approach for integrating AI cybersecurity with traditional cybersecurity principles

The push to implement AI security solutions in automotive systems responds to rapidly evolving threats and raises the need to secure AI systems in relation with the other components and services of the autonomous car. In particular, AI cybersecurity should be integrated with traditional cybersecurity principles.

Increasing dependence on AI for critical functions and services in AVs will not only create greater incentives for attackers to target those algorithms, but will also step up the potential for each successful attack to have more severe consequences [233]. Best practices for secure systems often ignore that an AV is a multidimensional environment with different components that may themselves include one or several AI models of different natures. In light of

this, ensuring cybersecurity in AVs requires an end-to-end holistic approach taking into account all the different components, the diversity of AI systems, and their interactions. Building security as an integrated procedure that involves various systems and takes into account all the phases of the AI systems is vital for the resilience of systems to potential security breaches. Applying defence-in-depth strategies plays a significant role in measuring and enforcing security requirements. Integrating AI cybersecurity, for all the steps of the AI lifecycle, with traditional security principles is very important, since a missing vulnerability may jeopardize the security of the whole autonomous vehicle.

Even if a system is designed and developed with security in mind, systems change over time with additional equipment, software and functionalities. This situation imposes the need for an integrated approach that needs to be maintained and updated, capturing all the systems, the AI models and their interactions. Towards this end, companies and organizations involved in the automotive sector need to update their cybersecurity policies accordingly.

There is a crucial need for a holistic approach integrating AI with traditional cybersecurity principles as well as a thorough documentation of AI systems in automotive context. Contrary to classical secure software development, for which "prepared statements", to avoid for example SQL injection attacks, are readily available, security patterns for the design and implementation of AI-based components are missing. This is nonetheless needed for the AI design and development: Securing AI pipelines throughout the whole AI lifecycle requires tamper-resistant implementations of each stage, mutual authentication between all the stages and confidentiality/integrity at the interfaces between the different stages. In the automotive sector, this translates into recommending tamper-resistant sensor, strongly authenticated components of the on-board network, adversarially trained on-board models, limited plasticity after deployment, etc.

RECOMMENDATIONS

- Establish security processes in the organizations integrating AI particularities.
- Promote security by design principles when it comes with deployment and development of AI in automotive context.
- Promote the use of standardised components and homogeneous AI solutions in automotive context.
- Ensure proper governance of AI cybersecurity policy in the organizations defining specific roles and responsibilities.
- Create an AI cybersecurity culture across the automotive ecosystem.
- Promote innovation and R&D activities for incorporating AI cybersecurity in the organizations.
- Promote dialogues between industrial actors to ensure interoperability in the development of AI solutions.
- Promote security patterns for the design and implementation of the AI-based components.
- Promote research projects on the security of AI components for autonomous driving.
- Implement solutions that can detect if not prevent the potential jamming of sensors.

4.4 Incident handling and vulnerability discovery related to AI and lessons learned

In many organisations, although cybersecurity teams know the main threats to which many components and systems in AVs are exposed, often people only become truly aware of the importance of security when they suffer an incident or discover a vulnerability. Despite the vast publicity regarding security vulnerabilities, the related awareness and commitment to security remains significantly low, especially with regards to vulnerabilities of AI systems.

It is worth highlighting that, in many cases, vulnerability discovery may influence the security practices more than the a-priori information about the existence of potential high risks. Optimistic bias is the main reason for this situation, stemming from the belief of many people that they are less likely to experience a negative event, because either they do not have the sufficient knowledge on actual risks or they are motivated to underestimate the risks. In this context, the optimistic bias may have a robust negative impact on the perception of AI security risks in the automotive.

The absence of AI security awareness and an inadequate AI security training also aggravate and perpetuate this bias. In this context, given the lack of information about the expected outcomes of a potential breach, the case studies and the first-hand accounts of security incidents and other security shortcomings will require significant amount of time to be properly resolved using current practices. It is

therefore essential to apply real-world training in order to deal with the negative AI security impacts of optimistic bias.

A clear and established cybersecurity incident handling and response plan should be considered, taking into account the increased number of digital components in the vehicle and in particular, the ones based on AI. An AI incident could be considered as an incident in which the behaviour of the vehicle as dictated by the planning module of the AV system is susceptible to cause harm, either due to an intentional malicious attack or due to the failure of an element in the ML pipeline. This may include potential violations of privacy and security, such as an external attacker attempting to manipulate the model or steal data encoded in the model, or incorrect predictions that can cause dangerous situations in which a traffic accident may happen.

A well-structured and domain-specific plan of action that immediately acts following an AI security breach or failure is essential in order to reduce the incident costs and damages to the organizations and the end users of the autonomous cars. There should also be a way to supervise AI systems and detect bad predictions (e.g. by comparing it against some ground truth such as maps and/or V2X messages from external sources). Frameworks further incorporating AI weaknesses (e.g. red-teaming) or penetration testing considering AI specific issues should also be considered.

The apparition of new actors with no previous experience in security incidents raises the need to build a cybersecurity culture to be able to comprehend the potential vulnerabilities and the underlying threats inherent to their systems, in order to know the correct steps to secure systems and to prioritize actions in case of incident. All stakeholders involved in the automotive supply chain should then stay aware of the growing AI threat landscape, in order to be able to map the risks and attacks to business operations. Lastly, processes for lessons learned when experiencing an AI security incident can also stimulate the creation of a security program across the whole supply chain.

RECOMMENDATIONS

- Adapt incident response plan to include AI particularities.
- Establish a culture of learning from AI security incidents.
- Promote knowledge sharing.
- Promote the use of mandatory standards for AI security incidents reporting.
- Organise disaster drills, involving high management, so that they understand the potential impact in case a vulnerability is discovered.
- Develop simulated incidents for raising awareness and knowledge in this sector.

4.5 Limited capacity and expertise on AI cybersecurity in the automotive industry

The absence of sufficient security knowledge and expertise among developers and system designers on AI cybersecurity is a major barrier that hampers the integration of security in the automotive sector. Many organisations associate security directly to the extent to which developers use security practices. The lack of AI knowledge by developers is then the source of several issues that may allow attackers to easily target AI components of AVs. First, the development processes do not include security tests and code analysis specific to AI components. Second, AI system designers tend to have a limited expertise on the domain of application, leading to poor design decisions.

As a result of these security shortcomings, AI security is often an afterthought, and AI security controls implemented as add-ons without full integration may arise, leading to complex, expensive, and hard to maintain architectures. This situation is fertile grounds for security vulnerabilities. Even if an AI system is designed and developed with security in mind, the volatility of AI systems, that need to be constantly updated with additional equipment, software and functionalities, imposes the need for AI security-aware security teams in order to authorize and track the changes as well as to evaluate potential AI security issues as part of a configuration management process.

Most people are not trained either properly or not at all in order to be able to recognize the security implications of AI software requirements. They do not know the security implications of the way that AI software is modelled, architected, designed, implemented, tested and prepared for distribution and deployment. Under these circumstances, AI software may not only deviate from its predefined security requirements but also these requirements may have been inadequate in the first place for its use in AVs. Without such knowledge, developers may not even recognize the security implications of certain design

and implementation choices, and that their mistakes and omissions in the development phase can lead to exploitable vulnerabilities in the software when it becomes operational.

The lack of project management to support and encourage the developers and designers to become AI security-aware through education and training is an aspect of paramount importance. Companies may also lack the resources to offer AI security training, developers may remain focused mainly on their primary functional task and security ignoring the AI cybersecurity features. Even if developers want to bear security in mind, sometimes there are budget and time related limitations posed by the top management and other stakeholders. It is then essential that project managers obtain an appropriate level of AI cybersecurity education and training in order to be less likely to make decisions that undermine the security of AVs.

Security awareness should not stop with the developers and managers, since it is important for all members associated with an AI software to receive security training to ensure that AI cybersecurity is a core concern in AVs. AI cybersecurity training of all involved parties will also help to make cybersecurity more prominent during discussion, planning and board meetings towards a secure automotive sector. In order to avoid damages that may deteriorate the reputation of the companies, the entire organisations must be aware of the importance of the implementation of AI security in AVs, and the consequences of not considering it as a priority objective. People should be trained and understand that cybersecurity of AI is not only about countering ML adversarial attacks, but also includes (together with adversarial mitigation measures) aspects from traditional cybersecurity, e.g. forensics, incident response, etc. In the automotive industry the AI systems should be designed by teams where automotive domain experts, ML experts and cybersecurity experts can collaborate.

RECOMMENDATIONS

- Integrate AI cybersecurity particularities in the whole organization policy.
- Create diverse teams consisted of experts from ML related fields, cybersecurity and the automotive sector.
- Involve mentors assisting the adoption of AI security practices in the organizations.
- Launch security education and training focused on AI systems cybersecurity and their integration across the automotive ecosystem.
- Deploy tools inside the continuous integration toolchain system that allows automated security testing of each pull request. This allows for an improved response and a better vulnerability remediation as they appear.
- Bring industry expertise to academic curriculum by welcoming lead people in the field to guest lectures or by defining special courses that tackle this topic.

REFERENCES

[1] 'Artificial Intelligence Cybersecurity Challenges - Threat Landscape for Artificial Intelligence', ENISA, 2020.

[2] United Nations - Economic Commission for Europe, *UN Regulation on uniform provisions concerning the approval of vehicles with regards to cyber security and cyber security management system*. .

[3] 'SAE J3016B: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles'. 2018.

[4] Y. Sun, D. Olaru, B. Smith, S. Greaves, and A. Collins, 'Road to autonomous vehicles in Australia: an exploratory literature review', *Road Transp. Res. J. Aust. N. Z. Res. Pract.*, vol. 26, no. 1, p. 34, 2017.

[5] T. Imai, 'Legal regulation of autonomous driving technology: Current conditions and issues in Japan', *IATSS Res.*, vol. 43, no. 4, pp. 263–267, 2019.

[6] Singapore Standards Council, 'Technical Reference 68 - Autonomous vehicles'. 2019.

[7] B. Clark, G. Parkhurst, and M. Ricci, 'Understanding the socioeconomic adoption scenarios for autonomous vehicles: A literature review', University of the West of England, Project Report, 2016.

[8] National Highway Traffic Safety Administration (NHTSA), 'Automated Driving Systems 2.0: A Vision for Safety', U.S. Department of Transportation, 2017.

[9] D. Ticoll, 'Automated Vehicles in Toronto', Innovation Policy Lab , Munk School of Global Affairs, University of Toronto, Discussion Paper.

[10] ERTRAC Working Group 'Connectivity and Automated Driving', 'Connected Automated Driving Roadmap', ERTRAC, 2019.

[11] C. Krupitzer, V. Lesch, M. Pfannemüller, C. Becker, and M. Segata, 'A Modular Simulation Framework for Analyzing Platooning Coordination', in *Proceedings of the 1st ACM MobiHoc Workshop on Technologies, mOdels, and Protocols for Cooperative Connected Cars*, New York, NY, USA, 2019, pp. 25–30.

[12] 'Cooperative, connected and automated mobility (CCAM)', *Mobility and Transport - European Commission*, 2016. [Online]. Available: https://ec.europa.eu/transport/themes/its/c-its_en. [Accessed: 01-Dec-2020].

[13] LSH Auto UK, *Mercedes-Benz MBUX Artificial Intelligence | LSH Auto UK*. 2019.

[14] 'Ad-Hoc Working Group on Artificial Intelligence Cybersecurity'. [Online]. Available: https://www.enisa.europa.eu/topics/iot-and-smart-infrastructures/artificial_intelligence/adhoc_wg_calls. [Accessed: 01-Dec-2020].

[15] Andy Greenberg, 'Hackers Remotely Kill a Jeep on the Highway—With Me in It', *Wired*, 2015.

[16] K. (Curtis) Zeng *et al.*, 'All Your {GPS} Are Belong To Us: Towards Stealthy Manipulation of Road Navigation Systems', in *Proceedings of the 27th {USENIX} Security Symposium*, 2018, pp. 1527–1544.

[17] J. Petit, B. Stottelaar, and M. Feiri, 'Remote Attacks on Automated Vehicles Sensors : Experiments on Camera and LiDAR', presented at the Black Hat Europe, 2015.

[18] H. Shin, D. Kim, Y. Kwon, and Y. Kim, 'Illusion and Dazzle: Adversarial Optical Channel Exploits Against Lidars for Automotive Applications', in *Proceedings of the International Conference on Cryptographic Hardware and Embedded Systems*, 2017, pp. 445–467.

[19] I. Foster, A. Prudhomme, K. Koscher, and S. Savage, 'Fast and vulnerable: A story of telematic failures', in *Proceedings of the 9th USENIX Workshop On Offensive Technologies (WOOT 15)*, Washington, D.C., 2015.

[20] K. Eykholt *et al.*, 'Robust Physical-World Attacks on Deep Learning Visual Classification', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.

[21] KPMG International, '2020 Autonomous Vehicles Readiness Index', 2020.

[22] European Commission, 'A European strategy on Cooperative Intelligent Transport Systems, a milestone towards cooperative, connected and automated mobility', COM/2016/0766 final, 2016.

[23] 'Platform: C-Roads'. [Online]. Available: https://www.c-roads.eu/platform.html. [Accessed: 07-Dec-2020].

[24] European Commission, 'On the road to automated mobility: An EU strategy for mobility of the future', COM(2018) 283 final, 2018.

[25] 'C-ITS Point of Contact'. [Online]. Available: https://cpoc.jrc.ec.europa.eu/index.html. [Accessed: 01-Dec-2020].

[26] 'European Commission Launches CCAM Single Platform - Connected Automated Driving Europe', 2019. [Online]. Available: https://connectedautomateddriving.eu/mediaroom/european-commission-launches-ccam-single-platform/. [Accessed: 01-Dec-2020].

[27] 'Expert group on cooperative, connected, automated and autonomous mobility (E03657)', 2019. [Online]. Available: https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupDetail&groupID=3657. [Accessed: 01-Dec-2020].

[28] 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', European Commission - Directorate-General for Research and Innovation, 2020.
[29] European Parliament and Council of the European Union, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. 2016.
[30] European Parliament and Council of the European Union, *Directive (EU) 2016/ 1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union*. .
[31] 'Canada's vehicle cyber security guidance', Transport Canada, 2020.
[32] ACEA, 'Principles of Automobile Cybersecurity', 2017.
[33] National Highway Traffic Safety Administration (NHTSA), 'Cybersecurity Best Practices for Modern Vehicles', U.S. Department of Transportation, 2016.
[34] M. Schaub and A. Zhao, 'China Releases Big Plan for Autonomous Vehicles', *China Law Insight*, 2020.
[35] Auto-ISAC, 'Best Practices - Executive Summary', 2016. [Online]. Available: https://automotiveisac.com/best-practices/. [Accessed: 01-Dec-2020].
[36] BSI, 'PAS 1885:2018 The fundamental principles of automotive cyber security. Specification'. 2018.
[37] BSI, 'PAS 11281:2018 Connected automotive ecosystems. Impact of security on safety. Code of practice'. 2018.
[38] ETSI, 'ETSI TS 102 940 - V1.3.1 - Intelligent Tranport Systems (ITS); Security; ITS communications security architecture and security management'. 2018.
[39] ETSI, 'TS 102 941 - V1.2.1 - Intelligent Transport Systems (ITS); Security; Trust and Privacy Management'. 2018.
[40] ETSI, 'TS 102 942 - V1.1.1 - Intelligent Transport Systems (ITS); Security; Access Control'. 2012.
[41] ETSI, 'TS 102 943 - V1.1.1 - Intelligent Transport Systems (ITS); Security; Confidentiality services'. 2012.
[42] 'ECTL - Documentation'. [Online]. Available: https://cpoc.jrc.ec.europa.eu/Documentation.html.
[43] SAE, 'SAE J3061: Cybersecurity Guidebook for Cyber-Physical Vehicle Systems'. 2016.
[44] 'ISO/SAE DIS 21434 - Road vehicles — Cybersecurity engineering'. 2020.
[45] 'SAE J3101: Hardware Protected Security for Ground Vehicles'. .
[46] ENISA, 'Good Practices For Security Of Smart Cars', ENISA, 2019.
[47] 'Cyber Security and Resilience of smart cars - Good practices and recommendations', ENISA, 2016.
[48] M. Craglia *et al.*, 'Artificial Intelligence–a European perspective', European Commission - Joint Research Centre, Science for policy, 2018.
[49] European Commission High Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI'. European Commission, 2019.
[50] S. D. Pendleton *et al.*, 'Perception, Planning, Control, and Coordination for Autonomous Vehicles', *Machines*, vol. 5, no. 1, p. 6, Mar. 2017.
[51] S. Edelstein, 'What is adaptive cruise control?', *Digital Trends*, 2020.
[52] Y. Song and C. Liao, 'Analysis and review of state-of-the-art automatic parking assist system', in *2016 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, 2016, pp. 1–6.
[53] J. Zhao, B. Liang, and Q. Chen, 'The key technology toward the self-driving car', *Int. J. Intell. Unmanned Syst.*, vol. 6, no. 1, pp. 2–20, Jan. 2018.
[54] C. Jensen, 'Are Blind Spots a Myth?', *Wheels Blog*, 2009. .
[55] 'Lane Change Assistance System : a lateral alert system', *Valeo*. [Online]. Available: https://www.valeo.com/en/lane-change-assistance-system/. [Accessed: 01-Dec-2020].
[56] A. Goodwin, 'Lane-keeping assist is preparing you for self-driving cars', *Roadshow*, 2017.
[57] K. Lim, Y. Hong, Y. Choi, and H. Byun, 'Real-time traffic sign recognition based on a general purpose GPU and deep-learning', *PLOS ONE*, vol. 12, no. 3, p. e0173317, 2017.
[58] Z. Wang, Y. Wu, and Q. Niu, 'Multi-Sensor Fusion in Automated Driving: A Survey', *IEEE Access*, vol. 8, pp. 2847–2868, 2020.
[59] G. Sharabok, 'Why Tesla Won't Use LIDAR', *Medium*, 2020. [Online]. Available: https://towardsdatascience.com/why-tesla-wont-use-lidar-57c325ae2ed5. [Accessed: 01-Dec-2020].
[60] A. J. Hawkins, 'Waymo will sell LIDAR to customers who won't compete with its robot taxi business', *The Verge*, 2019.

[61] V. De Silva, J. Roche, and A. Kondoz, 'Fusion of LiDAR and camera sensor data for environment sensing in driverless vehicles', Jan. 2018.

[62] S. Campbell *et al.*, 'Sensor Technology in Autonomous Vehicles : A review', in *2018 29th Irish Signals and Systems Conference (ISSC)*, Belfast, 2018, pp. 1–4.

[63] 'DITS - Data set of Italian Traffic Signs', 2016. [Online]. Available: http://users.diag.uniroma1.it/bloisi/ds/dits.html.

[64] *waymo-research/waymo-open-dataset*. Waymo Research, 2019.

[65] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*, 3rd ed. Pearson, 2009.

[66] J. Moor, 'The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years', *AI Mag.*, vol. 27, no. 4, p. 87, Dec. 2006.

[67] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[68] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, 'Gradient-based learning applied to document recognition', *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[69] A. Krizhevsky, I. Sutskever, and G. E. Hinton, 'ImageNet classification with deep convolutional neural networks', *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[70] K. Simonyan and A. Zisserman, 'Very Deep Convolutional Networks for Large-Scale Image Recognition', *ArXiv14091556 Cs*, Apr. 2015.

[71] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, 'Rethinking the Inception architecture for computer vision', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[72] C. Szegedy *et al.*, 'Going Deeper with Convolutions', preprint arxiv: 1409.4842, 2014.

[73] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[74] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, 'SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size', 2016.

[75] S. Ren, K. He, R. Girshick, and J. Sun, 'Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[76] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, 'You Only Look Once: Unified, Real-Time Object Detection', 2016.

[77] W. Ali, S. Abdelkarim, M. Zahran, M. Zidan, and A. E. Sallab, 'YOLO3D: End-to-end real-time 3D Oriented Object Bounding Box Detection from LiDAR Point Cloud', in *Proceedings of the European Conference on Computer Vision Workshop*, 2018.

[78] Y. Zhou and O. Tuzel, 'VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4490–4499.

[79] C. R. Qi, H. Su, M. Kaichun, and L. J. Guibas, 'PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 77–85.

[80] V. Badrinarayanan, A. Kendall, and R. Cipolla, 'SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation', in *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, vol. 39, pp. 2481–2495.

[81] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, 'ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation', *ArXiv160602147 Cs*, Jun. 2016.

[82] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, 'Pyramid Scene Parsing Network', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[83] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, 'Rethinking Atrous Convolution for Semantic Image Segmentation', 2017.

[84] K. He, G. Gkioxari, P. Dollar, and R. Girshick, 'Mask R-CNN', in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[85] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, 'A Survey of Deep Learning Techniques for Autonomous Driving', *J. Field Robot.*, 2019.

[86] D. Barnes, W. Maddern, G. Pascoe, and I. Posner, 'Driven to Distraction: Self-Supervised Distractor Learning for Robust Monocular Visual Odometry in Urban Environments', in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2018, pp. 1894–1900.

[87] I. A. Barsan, S. Wang, A. Pokrovsky, and R. Urtasun, 'Learning to Localize Using a LiDAR Intensity Map', in *Conference on Robot Learning*, 2018, pp. 605–616.

[88] W. Luo *et al.*, 'Multiple Object Tracking: A Literature Review', 2017.

[89] S. Agarwal, J. O. D. Terrail, and F. Jurie, 'Recent Advances in Object Detection in the Age of Deep Convolutional Neural Networks', *ArXiv180903193 Cs*, 2019.
[90] A. K. Ushani and R. M. Eustice, 'Feature Learning for Scene Flow Estimation from LIDAR', presented at the Conference on Robot Learning, 2018, pp. 283–292.
[91] W. Luo, B. Yang, and R. Urtasun, 'Fast and Furious: Real Time End-to-End 3D Detection, Tracking and Motion Forecasting with a Single Convolutional Net', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 3569–3577.
[92] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.
[93] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, 'Learning representations by back-propagating errors', *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
[94] S. Hochreiter and J. Schmidhuber, 'Long short-term memory', *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[95] M. Ghallab, D. Nau, and P. Traverso, *Automated Planning and Acting*. Cambridge University Press, 2016.
[96] M. Newman, *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010.
[97] J. Clark and D. A. Holton, *A first look at graph theory*. World Scientific, 1991.
[98] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. .
[99] V. Francois-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, *An Introduction to Deep Reinforcement Learning*. 2018.
[100] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser, 'Simultaneous Localization and Mapping: A Survey of Current Trends in Autonomous Driving', *IEEE Trans. Intell. Veh.*, vol. 2, no. 3, 2017.
[101] F. Leon and M. Gavrilescu, 'A Review of Tracking, Prediction and Decision Making Methods for Autonomous Driving', preprint arXiv: 1909.07707, 2019.
[102] A. Bar Hillel, R. Lerner, D. Levi, and G. Raz, 'Recent progress in road and lane detection: a survey', *Mach. Vis. Appl.*, vol. 25, no. 3, pp. 727–745, 2014.
[103] *[CVPR'20 Workshop on Scalability in Autonomous Driving] Keynote - Andrej Karpathy*. 2020.
[104] Y. Saadna and A. Behloul, 'An overview of traffic sign detection and classification methods', *Int. J. Multimed. Inf. Retr.*, vol. 6, no. 3, pp. 193–210, 2017.
[105] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, 'Traffic-Sign Detection and Classification in the Wild', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2110–2118.
[106] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, 'Detection of traffic signs in real-world images: The German traffic sign detection benchmark', in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 2013, pp. 1–8.
[107] H.-K. Kim, J. H. Park, and H.-Y. Jung, 'An Efficient Color Space for Deep-Learning Based Traffic Light Recognition', *J. Adv. Transp.*, vol. 2018, pp. 1–12, Dec. 2018.
[108] H.-K. Kim, K.-Y. Yoo, J. H. Park, and H.-Y. Jung, 'Traffic Light Recognition Based on Binary Semantic Segmentation Network', *Sensors*, vol. 19, no. 7, 2019.
[109] S. Alvarez, 'Tesla Autopilot's stop sign, traffic light recognition and response is operating in "Shadow Mode"', *TESLARATI*, 2019.
[110] Kai Wang, B. Babenko, and S. Belongie, 'End-to-end scene text recognition', in *2011 International Conference on Computer Vision*, 2011, pp. 1457–1464.
[111] Phuc Xuan Nguyen, K. Wang, and S. Belongie, 'Video text detection and recognition: Dataset and benchmark', in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2014, pp. 776–783.
[112] B. Shi, X. Bai, and C. Yao, 'An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, 2017.
[113] D. Karatzas *et al.*, 'ICDAR 2013 Robust Reading Competition', in *Proceedings of the 12th International Conference on Document Analysis and Recognition*, 2013, pp. 1484–1493.
[114] S. Reddy, M. Mathew, L. Gomez, M. Rusinol, D. Karatzas., and C. V. Jawahar, 'RoadText-1K: Text Detection & Recognition Dataset for Driving Videos', in *Proceedings of the International Conference on Robotics and Automation*, 2020.
[115] Q. Kong, Y. Cao, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, 'Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems', preprint arXiv: 1904.03476, 2019.
[116] M. K. Nandwana and T. Hasan, 'Towards Smart-Cars That Can Listen: Abnormal Acoustic Event Detection on the Road', presented at the Interspeech 2016, 2016, pp. 2968–2971.

[117] B. Paden, M. Čáp, S. Z. Yong, D. Yershov, and E. Frazzoli, 'A Survey of Motion Planning and Control Techniques for Self-Driving Urban Vehicles', *IEEE Trans. Intell. Veh.*, vol. 1, no. 1, pp. 33–55, Mar. 2016.
[118] J. Levinson, M. Montemerlo, and S. Thrun, 'Map-Based Precision Vehicle Localization in Urban Environments', *Robot. Sci. Syst.*, p. 8, 2007.
[119] S. Thrun, 'Probabilistic robotics', *Commun. ACM*, vol. 45, no. 3, Mar. 2002.
[120] L. A. Marina, B. Trasnea, T. Cocias, A. Vasilcoi, F. Moldoveanu, and S. M. Grigorescu, 'Deep Grid Net (DGN): A Deep Learning System for Real-Time Driving Context Understanding', in *Proceedings of the 3rd IEEE International Conference on Robotic Computing*, 2019, pp. 399–402.
[121] C. Seeger, A. Muller, L. Schwarz, and M. Manz, 'Towards Road Type Classification with Occupancy Grids', *IEEE Intell. Veh. Symp.*, p. 4, 2016.
[122] W. Schwarting, J. Alonso-Mora, and D. Rus, 'Planning and Decision-Making for Autonomous Vehicles', *Annu. Rev. Control Robot. Auton. Syst.*, vol. 1, no. 1, pp. 187–210, 2018.
[123] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, 'A Survey of Autonomous Driving: Common Practices and Emerging Technologies', *IEEE Access*, vol. 8, pp. 58443–58469, 2020.
[124] H. Bast *et al.*, 'Route Planning in Transportation Networks', in *Algorithm Engineering: Selected Results and Surveys*, L. Kliemann and P. Sanders, Eds. Cham: Springer International Publishing, 2016, pp. 19–80.
[125] Z. Zhuang, J. Wang, Q. Qi, H. Sun, and J. Liao, 'Toward Greater Intelligence in Route Planning: A Graph-Aware Deep Learning Approach', *IEEE Syst. J.*, vol. 14, no. 2, pp. 1658–1669, 2020.
[126] X. Zhou, M. Su, Z. Liu, Y. Hu, B. Sun, and G. Feng, 'Smart Tour Route Planning Algorithm Based on Naïve Bayes Interest Data Mining Machine Learning', *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 2, p. 112, Feb. 2020.
[127] M. Kuderer, S. Gulati, and W. Burgard, 'Learning driving styles for autonomous vehicles from demonstration', in *2015 IEEE International Conference on Robotics and Automation*, 2015, pp. 2641–2646.
[128] C. Miyajima and K. Takeda, 'Driver-Behavior Modeling Using On-Road Driving Data: A new application for behavior signal processing', *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 14–21, 2016.
[129] E. Yurtsever, C. Miyajima, S. Selpi, and K. Takeda, 'Driving signature extraction', in *Proceedings of the 3rd International Symposium on Future Active Safety Technology Toward zero traffic accidents*, 2015.
[130] E. Yurtsever, C. Miyajima, and K. Takeda, 'A Traffic Flow Simulation Framework for Learning Driver Heterogeneity from Naturalistic Driving Data using Autoencoders', *Int. J. Automot. Eng.*, vol. 10, no. 1, pp. 86–93, 2019.
[131] B. R. Kiran *et al.*, 'Deep Reinforcement Learning for Autonomous Driving: A Survey', preprint arXiv: 2002.00444, 2020.
[132] L. Sun, C. Peng, W. Zhan, and M. Tomizuka, 'A Fast Integrated Planning and Control Framework for Autonomous Driving via Imitation Learning', presented at the ASME 2018 Dynamic Systems and Control Conference, 2018.
[133] M. Bansal, A. Krizhevsky, and A. Ogale, 'ChauffeurNet: Learning to Drive by Imitating the Best and Synthesizing the Worst', 2018.
[134] A. Kendall *et al.*, 'Learning to Drive in a Day', in *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2019, pp. 8248–8254.
[135] X. Pan, Y. You, Z. Wang, and C. Lu, 'Virtual to Real Reinforcement Learning for Autonomous Driving', preprint arXiv: 1704.03952, 2017.
[136] D. Barnes, W. Maddern, and I. Posner, 'Find your own way: Weakly-supervised segmentation of path proposals for urban autonomy', in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2017, pp. 203–210.
[137] L. Caltagirone, M. Bellone, L. Svensson, and M. Wahde, 'LIDAR-based driving path generation using fully convolutional neural networks', in *Proceedings of the 20th IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2017, pp. 1–6.
[138] S. H. Park, B. Kim, C. M. Kang, C. C. Chung, and J. W. Choi, 'Sequence-to-Sequence Prediction of Vehicle Trajectory via LSTM Encoder-Decoder Architecture', in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2018, pp. 1672–1678.
[139] D. Isele, R. Rahimi, A. Cosgun, K. Subramanian, and K. Fujimura, 'Navigating Occluded Intersections with Autonomous Vehicles Using Deep Reinforcement Learning', in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2018, pp. 2034–2039.
[140] K. K. Tan, Q.-G. Wang, and C. C. Hang, *Advances in PID Control*. Springer Science & Business Media, 2012.
[141] M. Morari, C. E. Garcia, and D. M. Prett, 'Model predictive control: Theory and practice', *IFAC Proc. Vol.*, vol. 21, no. 4, pp. 1–12, 1988.
[142] J. R. Quain and 2015, 'BMW 7 Series Tested: Here's The Tech You Get for $100K', *Tom's Guide*. [Online]. Available: https://www.tomsguide.com/us/bmw-7-series-tested,review-3040.html. [Accessed: 01-Dec-2020].

[143] Toyota Motor Corporation, 'Toyota's New "LQ" Wants to Build an Emotional Bond with Its Driver'. 2019.
[144] M. Bojarski *et al.*, 'End to End Learning for Self-Driving Cars', 2016.
[145] T. P. Lillicrap *et al.*, 'Continuous control with deep reinforcement learning', 2019.
[146] H. Xu, Y. Gao, F. Yu, and T. Darrell, 'End-To-End Learning of Driving Models From Large-Scale Video Datasets', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2174–2182.
[147] B. Wymann, C. Dimitrakakis, A. Sumner, E. Espie, and C. Guionneau, 'TORCS: The open racing car simulator', p. 5, 2015.
[148] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, 'CARLA: An Open Urban Driving Simulator', preprint arXiv: 1711.03938, 2017.
[149] S. Shah, D. Dey, C. Lovett, and A. Kapoor, 'Aerial Informatics and Robotics Platform', Technical Report.
[150] G. Rong *et al.*, 'LGSVL Simulator: A High Fidelity Simulator for Autonomous Driving', preprint arxiv: 2005.03778, 2020.
[151] S. Manivasagam *et al.*, 'LiDARsim: Realistic LiDAR Simulation by Leveraging the Real World', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
[152] L. Fridman, J. Terwilliger, and B. Jenik, 'DeepTraffic: Crowdsourced Hyperparameter Tuning of Deep Reinforcement Learning Systems for Multi-Agent Dense Traffic Navigation', presented at the Neural Information Processing Systems (NIPS 2018) Deep Reinforcement Learning Workshop, 2019.
[153] Ram Shankar Siva Kumar, David O'Brien, Jeffrey Snover, Kendra Albert, and Salome Viljoen, 'Failure Modes in Machine Learning - Security documentation'. [Online]. Available: https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning. [Accessed: 26-May-2020].
[154] N. Dalvi, P. Domingos, S. Sanghai, D. Verma, and others, 'Adversarial classification', in *Proceedings of the 10th ACM international conference on Knowledge discovery and data mining*, 2004, pp. 99–108.
[155] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. Tygar, 'Adversarial machine learning', in *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, 2011, pp. 43–58.
[156] P. A. Johnson, B. Tan, and S. Schuckers, 'Multimodal fusion vulnerability to non-zero effort (spoof) imposters', in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, 2010, pp. 1–5.
[157] A. Adler, 'Vulnerabilities in Biometric Encryption Systems', in *Proceedings of the International Conference on Audio- and Video-Based Biometric Person Authentication*, Berlin, Heidelberg, 2005, pp. 1100–1109.
[158] P. Fogla, M. Sharif, R. Perdisci, O. Kolesnikov, and W. Lee, 'Polymorphic Blending Attacks', presented at the {USENIX} Security Symposium, 2006, p. 16.
[159] C.-H. Huang, T.-H. Lee, L. Chang, J.-R. Lin, and G. Horng, 'Adversarial Attacks on SDN-Based Deep Learning IDS System', in *Proceedings of the International Conference on Mobile and Wireless Technology (ICMWT)*, Singapore, 2019, pp. 181–191.
[160] D. Lowd, 'Good word attacks on statistical spam filters', in *Proceedings of the 2nd Conference on Email and Anti-Spam*, 2005.
[161] A. Kołcz and C. H. Teo, 'Feature Weighting for Improved Classifier Robustness', in *Proceedings of the 6th Conference on Email and Anti-Spam*, 2009, p. 8.
[162] B. Biggio, B. Nelson, and P. Laskov, 'Poisoning attacks against support vector machines', in *Proceedings of the 29th International Conference on International Conference on Machine Learning*, 2012, pp. 1467–1474.
[163] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, 'A General Framework for Adversarial Examples with Objectives', *ACM Trans. Priv. Secur.*, vol. 22, no. 3, pp. 16:1–16:30, 2019.
[164] N. Carlini *et al.*, 'Hidden Voice Commands', in *Proceedings of the 25th {USENIX} Security Symposium*, 2016, pp. 513–530.
[165] R. Jia and P. Liang, 'Adversarial Examples for Evaluating Reading Comprehension Systems', in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.
[166] I. J. Goodfellow, J. Shlens, and C. Szegedy, 'Explaining and harnessing adversarial examples', in *Proceedings of the International conference on learning representations*, 2015.
[167] C. Szegedy *et al.*, 'Intriguing properties of neural networks', in *Proceedings of the International Conference on Learning Representations*, 2014.
[168] M. Großhans, C. Sawade, M. Brückner, and T. Scheffer, 'Bayesian Games for Adversarial Regression Problems', in *Proceedings of the International Conference on Machine Learning*, 2013, pp. 55--63.
[169] C. Liu, B. Li, Y. Vorobeychik, and A. Oprea, 'Robust Linear Regression Against Training Data Poisoning', in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, New York, NY, USA, 2017, pp. 91–102.

[170] B. Biggio, G. Fumera, and F. Roli, 'Multiple Classifier Systems under Attack', in *Proceedings of the International Workshop on Multiple Classifier Systems*, Berlin, Heidelberg, 2010, pp. 74–83.
[171] B. Biggio, I. Pillai, S. Rota Bulò, D. Ariu, M. Pelillo, and F. Roli, 'Is data clustering in adversarial settings secure?', in *Proceedings of the ACM workshop on Artificial intelligence and security*, New York, NY, USA, 2013, pp. 87–98.
[172] B. Biggio *et al.*, 'Poisoning Complete-Linkage Hierarchical Clustering', in *Proceedings of the Workshop on Syntactic Pattern Recognition*, Berlin, Heidelberg, 2014, pp. 42–52.
[173] J. G. Dutrisac and D. B. Skillicorn, 'Hiding clusters in adversarial settings', in *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*, 2008, pp. 185–187.
[174] N. Narodytska and S. Kasiviswanathan, 'Simple Black-Box Adversarial Attacks on Deep Neural Networks', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1310–1318.
[175] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, 'DeepFool: a simple and accurate method to fool deep neural networks', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2574–2582.
[176] L. Muñoz-González *et al.*, 'Towards Poisoning of Deep Learning Algorithms with Back-gradient Optimization', in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, New York, NY, USA, 2017, pp. 27–38.
[177] N. Papernot, P. McDaniel, A. Swami, and R. Harang, 'Crafting adversarial input sequences for recurrent neural networks', in *Proceedings of the IEEE Military Communications Conference*, 2016, pp. 49–54.
[178] T. Gu, B. Dolan-Gavitt, and S. Garg, 'BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain', preprint arXiv: 1708.06733, 2019.
[179] W. Uther and M. Veloso, 'Adversarial Reinforcement Learning', preprint, 1997.
[180] Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu, and M. Sun, 'Tactics of adversarial attack on deep reinforcement learning agents', in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, Melbourne, Australia, 2017, pp. 3756–3762.
[181] C. Xiao *et al.*, 'Characterizing Attacks on Deep Reinforcement Learning', preprint arxiv: 1907.09470, 2019.
[182] V. Behzadan and A. Munir, 'Vulnerability of Deep Reinforcement Learning to Policy Induction Attacks', in *Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition*, 2017, pp. 262–275.
[183] S. Li *et al.*, 'Stealthy Adversarial Perturbations Against Real-Time Video Classification Systems', in *Proceedings of the Network and Distributed Systems Security Symposium*, 2019, p. 15.
[184] A. Kurakin, I. J. Goodfellow, and S. Bengio, 'Adversarial machine learning at scale', in *Proceedings of the International Conference on Learning Representations*, 2016.
[185] O. Russakovsky *et al.*, 'ImageNet Large Scale Visual Recognition Challenge', *Int. J. Comput. Vis. IJCV*, vol. 115, no. 3, pp. 211–252, 2015.
[186] S. Qiu, Q. Liu, S. Zhou, and C. Wu, 'Review of Artificial Intelligence Adversarial Attack and Defense Technologies', *Appl. Sci.*, vol. 9, p. 909, 2019.
[187] E. Tabassi, K. J. Burns, M. Hadjimichael, A. D. Molina-Markham, and J. T. Sexton, 'A taxonomy and terminology of adversarial machine learning', Draft NISTIR 8269, 2019.
[188] B. Biggio and F. Roli, 'Wild patterns: Ten years after the rise of adversarial machine learning', *Pattern Recognit.*, vol. 84, pp. 317–331, 2018.
[189] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, 'Towards deep learning models resistant to adversarial attacks', in *International conference on learning representations*, 2018.
[190] S. Bubeck, 'Convex Optimization: Algorithms and Complexity', *Found. Trends® Mach. Learn.*, vol. 8, no. 3–4, pp. 231–357, 2015.
[191] I. J. Goodfellow, J. Shlens, and C. Szegedy, 'Explaining and harnessing adversarial examples', in *Proceedings of the International conference on learning representations*, 2015.
[192] Y. Dong *et al.*, 'Boosting Adversarial Attacks With Momentum', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9185–9193.
[193] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, 'Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, 2019.
[194] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, 'The limitations of deep learning in adversarial settings', in *Proceedings of IEEE European Symposium on Security and Privacy*, 2016, pp. 372–387.

[195] N. Carlini and D. Wagner, 'Towards evaluating the robustness of neural networks', in *Proceedings of the IEEE Symposium on Security and Privacy*, 2017.
[196] J. Su, D. V. Vargas, and K. Sakurai, 'One pixel attack for fooling deep neural networks', *IEEE Trans. Evol. Comput.*, 2019.
[197] M. M. Cisse, Y. Adi, N. Neverova, and J. Keshet, 'Houdini: Fooling Deep Structured Visual and Speech Recognition Models with Adversarial Examples', in *Proceedings of the Advances in Neural Information Processing Systems*, 2017, vol. 30, pp. 6977–6987.
[198] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, 'ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models', in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, New York, NY, USA, 2017, pp. 15–26.
[199] Z. Liu, P. Luo, X. Wang, and X. Tang, 'Deep Learning Face Attributes in the Wild', in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738.
[200] Y. Song, R. Shu, N. Kushman, and S. Ermon, 'Constructing Unrestricted Adversarial Examples with Generative Models', in *Proceedings of the Advances in Neural Information Processing Systems*, 2018, vol. 31, pp. 8312–8323.
[201] T. Hwang, 'Deepfakes - A Grounded Threat', CSET, 2020.
[202] Y. Cao *et al.*, 'Adversarial Sensor Attack on LiDAR-based Perception in Autonomous Driving', in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, New York, NY, USA, 2019, pp. 2267–2281.
[203] Y. Cao *et al.*, 'Adversarial Objects Against LiDAR-Based Autonomous Driving Systems', preprint arXiv: 1907.05418, 2019.
[204] J. Sun, Y. Cao, Q. A. Chen, and Z. M. Mao, 'Towards Robust LiDAR-based Perception in Autonomous Driving: General Black-box Adversarial Sensor Attack and Countermeasures', presented at the 29th {USENIX} Security Symposium ({USENIX} Security 20), 2020, pp. 877–894.
[205] C. Yan, W. Xu, and J. Liu, 'Can You Trust Autonomous Vehicles: Contactless Attacks against Sensors of Self-driving Vehicle', presented at the DEF CON, 2016, vol. 24.
[206] W. Xu, C. Yan, W. Jia, X. Ji, and J. Liu, 'Analyzing and Enhancing the Security of Ultrasonic Sensors for Autonomous Vehicles', *IEEE Internet Things J.*, vol. 5, no. 6, pp. 5015–5029, 2018.
[207] A. Hamdi, S. Rojas, A. Thabet, and B. Ghanem, 'AdvPC: Transferable Adversarial Perturbations on 3D Point Clouds', in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 241–257.
[208] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, 'A General Framework for Adversarial Examples with Objectives', *ACM Trans. Priv. Secur.*, vol. 22, no. 3, pp. 16:1–16:30, 2019.
[209] K. Eykholt *et al.*, 'Robust Physical-World Attacks on Deep Learning Visual Classification', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.
[210] Y. Luo, X. Boix, G. Roig, T. Poggio, and Q. Zhao, 'Foveation-based Mechanisms Alleviate Adversarial Examples', preprint arXiv: 1511.06292, 2016.
[211] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, 'Synthesizing robust adversarial examples', in *Proceedings of the International Conference on Machine Learning*, 2018, pp. 284–293.
[212] C. Berghoff, M. Neu, and A. von Twickel, 'Vulnerabilities of Connectionist AI Applications: Evaluation and Defense', *Front. Big Data*, vol. 3, 2020.
[213] R. Huang, B. Xu, D. Schuurmans, and C. Szepesvari, 'Learning with a Strong Adversary', preprint arXiv: 1511.03034, 2016.
[214] H. Hosseini, Y. Chen, S. Kannan, B. Zhang, and R. Poovendran, 'Blocking Transferability of Adversarial Examples in Black-Box Learning Systems', preprint arXiv: 1703.04318, 2017.
[215] B. Biggio *et al.*, 'One-and-a-Half-Class Multiple Classifier Systems for Secure Learning Against Evasion Attacks at Test Time', in *Proceedings of the International Workshop on Multiple Classifier Systems*, 2015, pp. 168–180.
[216] D. Meng and H. Chen, 'MagNet: A Two-Pronged Defense against Adversarial Examples', in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, New York, NY, USA, 2017, pp. 135–147.
[217] P. Samangouei, M. Kabkab, and R. Chellappa, 'Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models', in *Proceedings of the International Conference on Learning Representations*, 2018.
[218] C. Lyu, K. Huang, and H.-N. Liang, 'A Unified Gradient Regularization Family for Adversarial Examples', in *Proceedings of the IEEE International Conference on Data Mining*, 2015, pp. 301–309.
[219] Q. Zhao and L. D. Griffin, 'Suppressing the Unusual: towards Robust CNNs using Symmetric Activation Functions', preprint arXiv: 1603.05145, 2016.

[220] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, 'Distillation as a defense to adversarial perturbations against deep neural networks', in *2016 IEEE Symposium on Security and Privacy*, 2016, pp. 582–597.
[221] C. Qin *et al.*, 'Verification of non-linear specifications for neural networks', in *Proceedings of the international conference on learning representations (ICLR)*, 2019.
[222] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, 'Certified Adversarial Robustness via Randomized Smoothing', *ArXiv190202918 Cs Stat*, 2019.
[223] P. Wild, P. Radu, L. Chen, and J. Ferryman, 'Robust multimodal face and fingerprint fusion in the presence of spoofing attacks', *Pattern Recognit.*, vol. 50, pp. 17–25, 2016.
[224] B. Biggio, G. Fumera, G. L. Marcialis, and F. Roli, 'Statistical Meta-Analysis of Presentation Attacks for Secure Multibiometric Systems', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 561–575, 2017.
[225] C. Sitawarin, A. Nitin Bhagoji, A. Mosenia, M. Chiang, and P. Mittal, 'DARTS: Deceiving Autonomous Cars with Toxic Signs', preprint arxiv: 1802.06430, 2018.
[226] N. Morgulis, A. Kreines, S. Mendelowitz, and Y. Weisglass, 'Fooling a Real Car with Adversarial Traffic Signs', preprint arxiv: 1907.00374, 2019.
[227] S. Povolny and S. Trivedi, 'Model Hacking ADAS to Pave Safer Roads for Autonomous Vehicles', *McAfee Blogs*, 2020.
[228] D. Nassi, R. Ben-Netanel, Y. Elovici, and B. Nassi, 'MobilBye: Attacking ADAS with Camera Spoofing', preprint arxiv: 1906.09765, 2019.
[229] Tencent Keen Security Lab, 'Experimental security research of Tesla autopilot', 2019.
[230] A. Chernikova, A. Oprea, C. Nita-Rotaru, and B. Kim, 'Are Self-Driving Cars Secure? Evasion Attacks Against Deep Neural Networks for Steering Angle Prediction', in *Proceedings of the IEEE Security and Privacy Workshops*, 2019, pp. 132–137.
[231] G. Jocher *et al.*, *ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements*. Ultralytics LLC, 2020.
[232] H. Kume, 'Tesla teardown finds electronics 6 years ahead of Toyota and VW', *Nikkei Asia*, 2020.
[233] J. Wolff, 'How to improve cybersecurity for artificial intelligence', 2020.
[234] M. Alonso Raposo *et al., 'The future of road transport'*, EUR 29748 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-14318-5, doi:10.2760/668964, 2019.
[235] I. Nai Fovino *et al.*, '*Cybersecurity, our digital anchor*', EUR 30276 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-19957-1, doi:10.2760/352218, 2020.

KJ-NA-30568-EN-N

ABOUT ENISA

The European Union Agency for Network and Information Security (ENISA) is a centre of network and information security expertise for the EU, its member states, the private sector and Europe's citizens. ENISA works with these groups to develop advice and recommendations on good practice in information security. It assists EU member states in implementing relevant EU legislation and works to improve the resilience of Europe's critical information infrastructure and networks. ENISA seeks to enhance existing expertise in EU member states by supporting the development of cross-border communities committed to improving network and information security throughout the EU. More information about ENISA and its work can be found at www.enisa.europa.eu.

ABOUT JRC

The Joint Research Centre is the European Commission's science and knowledge service. The JRC is a Directorate-General of the European Commission. Our researchers provide EU and national authorities with solid facts and independent support to help tackle the big challenges facing our societies today. Our headquarters are in Brussels and we have research sites in five Member States: Geel (Belgium), Ispra (Italy), Karlsruhe (Germany), Petten (the Netherlands) and Seville (Spain). Our work is largely funded by the EU's budget for Research and Innovation. We create, manage and make sense of knowledge, delivering the best scientific evidence and innovative tools for the policies that matter to citizens, businesses and governments.

For more information, visit https://ec.europa.eu/jrc.

enisa.europa.eu

ec.europa.eu/jrc

Publications Office
of the European Union

doi:10.2760/551271
ISBN 978-92-76-28646-2