



D6.3 Lessons learned and policy recommendations

Dissemination level	Public (PU)
Work package	WP6
Task:	T6.5
Deliverable lead:	Rupprecht Consult
Version	V1.0
Submission date	28/10/2025
Due date	31/10/2025

Authors

Authors in alphabetical order		
Name	Organisation	Email
Anton Wijbenga	MAPTM	anton.wijbenga@maptm.nl
Guido Linden	IKA	guido.linden@ika.rwth-aachen.de
Jos den Ouden	TUE	j.h.v.d.ouden@tue.nl
Justin Hidalgo	IDIADA	justin.hidalgo@idiada.com
Lakshya Pandit	RUPPRECHT CONSULT	l.pandit@rupprecht-consult.eu
Paola Natalia Cañas	VICOMTECH	pncanas@vicomtech.org
Till Beemelmans	IKA	till.beemelmans@ika.rwth-aachen.de

Control sheet

Version history			
Version	Date	Modified by	Summary of changes
0.1	16/05/2025	IDIADA	Literature study
0.2	25.06.2025	RUPPRECHT CONSULT	Restructured format for project outputs and policy recommendations
0.3	18.09.2025	RUPPRECHT CONSULT, IKA, VICOMTECH, TTTA, IFAG, SIE-NL, VIF, and IDIADA	Inputs on ethical evaluation, data, and XAI + Inputs on UC insights and recommendations + Inputs on tools and testing facilities
0.4	16.10.2025	RUPPRECHT CONSULT	Inputs for text and visuals in Chapters 1, 8 and 9 + references
0.5	17.10.2025	TUE	Input to UC2.1
1.0	24.10.2025	RUPPRECHT CONSULT, MAPTM, and VICOMTECH	Final inputs based on review feedback

Peer review		
	Reviewer name	Date
Reviewer 1.1	Guido Linden (for Ch. 1, 2, 6, 7.4, 8, 9); ika	20/10/2025
Reviewer 1.2	Bastian Lampe (for Ch. 7.1 and 7.3); ika	20/10/2025
Reviewer 2	Mónica Díaz de Mendivil (VICOMTECH)	21/10/2025



**Funded by
the European Union**

Project funded by



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs,
Education and Research EAER
**State Secretariat for Education,
Research and Innovation SERI**

Disclaimer

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

This work has received funding from the Swiss State Secretariat for Education, Research, and Innovation (SERI).

TABLE OF CONTENTS

EXECUTIVE SUMMARY.....	6
LIST OF FIGURES	7
LIST OF ACRONYMS.....	8
1. INTRODUCTION.....	9
1.1. Project overview and relevance of policy recommendations.....	9
1.2. Objectives.....	10
1.3. Target audience.....	10
1.4. Structure of the deliverable	10
2. AI IMPACT AND CHALLENGES	12
2.1 Preface	12
2.2 AI impact on regulations and policies.....	13
2.3 AI impact on NATM	15
2.4 AI impact on safety assurance standards	16
2.5 Challenges	18
3. ETHICAL EVALUATION AND USER PERSPECTIVE	20
3.1 Ethical evaluation framework for AI in CCAM	20
3.2 User needs, expectations and concerns	23
4. DATA - Life Cycle Management and Generation	25
4.1 ML DevOps Data governance and provenance mechanisms.....	25
4.2 Privacy Preserving ML	26
5. XAI - Explainable Continuous Model Development.....	28
5.1 ML Development and life cycle management framework.....	28
5.2 AI algorithm development.....	29
6. TOOLS AND TESTING FACILITIES	32
6.1 Physical infrastructure and vehicle setup	32
6.2 Virtual tools and testing.....	33
7. DEPLOY & TEST - AI-based CCAM Deployment.....	34
7.1. UC1: Trustworthy Perception Systems for CCAM	34
7.2. UC2: AI-extended situational awareness/understanding	35
7.3. UC3: Trustworthy decision making	38
7.4. UC4: AI in Traffic Management.....	40

8.	POLICY RECOMMENDATIONS	43
8.1.	Human centred ethical evaluation	43
8.2.	Data and life-cycle management	44
8.3.	Explainable Continuous Model Development	45
8.4.	Tools, Testing, Deployment & Validation	46
9.	CONCLUSION	48
	REFERENCES	49

EXECUTIVE SUMMARY

This report provides a comprehensive examination of the integration of Trustworthy Artificial Intelligence within Cooperative, Connected, and Automated Mobility (CCAM) systems, addressing both technical and regulatory dimensions. Chapters 1 to 7 lay the foundation for understanding current challenges, methodologies, and the strategic roadmap for adoption.

The introductory chapter establishes the relevance of AI in CCAM, setting forth the context and motivations for the study. It underscores the necessity of transparency and trust in AI-driven mobility solutions, particularly as these systems interact with a wide range of stakeholders from developers to end-users.

Subsequent chapters detail the legislative landscape governing AI and automotive technologies in the European Union, with special attention to the evolving EU Vehicle Regulations Framework and the proposed EU AI Act. These frameworks are analysed to understand their implications for the deployment and certification of XAI features within CCAM platforms. The human-centric methodological approach is outlined, describing the systematic assessment of developed AI solutions, best practices, and testing protocols. Data collection strategies and evaluation benchmarks are presented to ensure robust and reproducible outcomes.

A significant portion of the analysis is dedicated to the role of XAI tools and techniques, highlighting their importance in making AI decisions interpretable and actionable for system developers, regulators, and users alike. The report also examines case studies and pilot deployments that demonstrate the practical impact and added value of XAI in mobility solutions. Chapters 6 and 7 synthesise these findings, discussing both the current state of adoption and the gaps that remain. The discussion extends to the technological, procedural, and organisational barriers that must be overcome to ensure widespread and effective implementation of XAI in CCAM systems. Chapter 8 summarises the in-depth recommendations from previous sections into policy recommendations with visualisations.

This report provides readers with key insights and lessons learned along with recommendations based on the solutions and tools developed in AITHENA.

LIST OF FIGURES

Figure 1: AITHENA's approach towards XAI inclusion in CCAM system	9
Figure 2: EU Vehicle Regulations Framework with EU AI Act.....	14
Figure 3: Image showing how developers of AI systems can communicate relevant information about the system to different users via Model Cards (Source: AITHENA 2023).....	29
Figure 4: Recommendations for human-centric ethical evaluation.....	43
Figure 5: Recommendations for data and life-cycle management.....	44
Figure 6: Recommendations for Explainable Continuous Model Development	45
Figure 7: Recommendations for Tools, Testing, Deployment, and Validation.....	46

LIST OF TABLES

Table 1: Comparison of Key Automotive Safety Standards in Addressing AI-Related Risks and Limitations	17
Table 2: Challenges and approaches to build trustworthy and interpretable AI systems.....	18

LIST OF ACRONYMS

ACRONYM	MEANING
ADAS	Advanced Driver-Assistance Systems
ADS	Automated Driving Systems
AI	Artificial Intelligence
AV	Autonomous Vehicle
CCAM	Connected, Cooperative, and Automated Mobility
DAPT	Domain-Adaptive Pre-Training
DL	Deep Learning
DPO	Data Protection Officer
DynViL	Dynamic Vehicle-in-the-Loop
EU	European Union
FL	Federated Learning
GDPR	General Data Protection Regulation
HMI	Human-Machine Interface
ICT	Information and Communication Technology
ISO	International Organization for Standardization
KPI	Key Performance Indicator
ML	Machine Learning
MLOps	Machine Learning Operations
MVM	Masked Video Modelling
NATM	New Assessment Test Method
OEM	Original Equipment Manufacturer
R&D	Research and Development
SOP	Standard Operating Procedure
TAD	Traffic Anomaly Detection
TRL	Technology Readiness Level
UNECE	United Nations Economic Commission for Europe
VideoViTs	Video Vision Transformers
VLM	Vision Language Models
V2X	Vehicle-to-Everything
WP	Work Package
XAI	Explainable Artificial Intelligence
XiL	X-in-the-Loop

1. INTRODUCTION

As artificial intelligence (AI) continues to transform the landscape of Connected, Cooperative, and Automated Mobility (CCAM), the need for robust, forward-thinking policy frameworks has never been more urgent. Navigating the complexities of explainability, ethical governance, and regulatory clarity is essential to harnessing AI's transformative potential while safeguarding public trust and safety. This chapter draws from the collective expertise and research advancements of the AITHENA project, and provides introduction to the objectives, target audience and structure for this deliverable.

1.1. Project overview and relevance of policy recommendations

The rapid advancement of AI in CCAM requires comprehensive policy frameworks to ensure safe, ethical, and accountable deployment. Current regulatory gaps leave significant uncertainties regarding liability, safety standards, and operational boundaries. Clear guidelines on explainability, privacy preservation, and ethical AI behaviour are crucial for widespread adoption and social acceptance. This deliverable summarises the main project results, outputs, and tools in an easy-to-read and concise format, as well as make policy recommendations for the exploitation of AI solutions in CCAM.

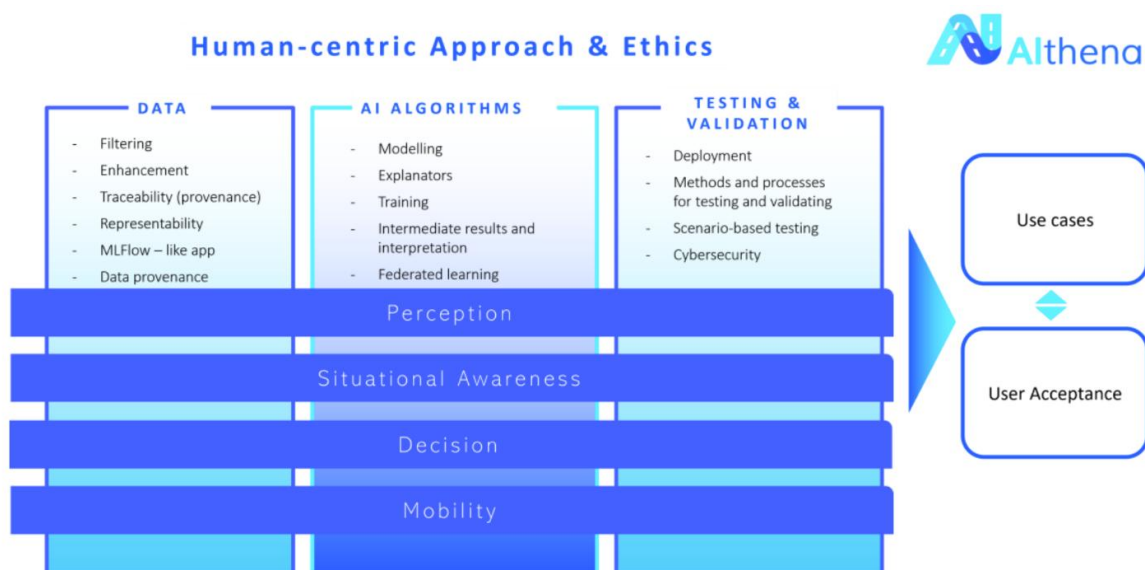


Figure 1: AITHENA's approach towards XAI inclusion in CCAM system

AI's challenges in explainability, privacy, ethics, and accountability hinder its full societal potential. It is this reason that AITHENA project contributes to build explainable AI (XAI) in CCAM development and testing frameworks, researching three main AI pillars (refer Figure 1) with a human-centric approach: data (real/synthetic data management), AI Algorithms (data fusion, hybrid AI approaches), and testing and validation (physical/virtual XiL set-ups with scalable MLOps). AITHENA proposes a set of Key Performance Indicators (KPIs) for XAI and analyses trade-offs between these dimensions, bridging the gap between advanced autonomous driving capabilities and transparency in AI decision-making.

While AITHENA's policy recommendations aim to connect technical progress with regulatory needs, the hope is to lay a foundation for more consistent XAI standards in CCAM, encourage cross-border cooperation in AI-enabled mobility, and offer useful guidance to authorities and stakeholders. By sharing findings and key insights, AITHENA aspires to contribute meaningfully to trustworthy AI governance in the mobility sector, supporting broader efforts to establish reliable standards and best practices.

1.2. Objectives

This deliverable is part of WP6 titled 'Impact: Exploitation, Dissemination and Standardisation'. The main objective of this deliverable is summarising the main project results, outputs, as well as make policy recommendations for the exploitation of AI solutions in CCAM while analysing the type of approval framework related to using AI in vehicles.

1.3. Target audience

This deliverable will be disseminated publicly informing all interested stakeholders about key insights and lessons learned on AITHENA's human centric AI solutions and its exploitation in CCAM followed by policy recommendations.

- The primary audience for this deliverable is policymakers who are responsible for developing regulatory frameworks for AI in mobility. For this audience, the policy recommendations provide comprehensive guidance on how to balance innovation with safety, privacy, and ethical considerations when creating legislation and policies for AI-enabled CCAM systems.
- A second audience is regulatory authorities (such as national transport agencies, vehicle certification bodies, and data protection authorities) who need to implement and enforce AI regulations in practice. For this audience, the deliverable offers practical frameworks for assessment, approval processes, and compliance monitoring of AI solutions in connected and automated mobility.
- The third audience is automotive industry stakeholders (including OEMs, Tier 1 suppliers, technology providers, and systems integrators) who are developing and deploying AI solutions in CCAM systems. For this audience, the recommendations provide clear guidance on regulatory requirements, technical standards, and best practices to ensure their AI solutions meet trustworthiness criteria and market acceptance.

1.4. Structure of the deliverable

The structure of this report is organised into distinct chapters, each addressing a key aspect of AITHENA's objectives and outputs.

- The introductory chapter provides an overview of the project's aims, contextualising the need for trustworthy and human-centric AI solutions in Connected, Cooperative and Automated Mobility (CCAM).

- Subsequent sections delve into the impact and challenges of AI adoption in the automotive sector, discussing both the technological foundations, and the regulatory challenges including type-approval, exploring different associated standards.
- Following the introduction and context-setting, the report details specific project results, outputs, highlighting key insights and lessons learned followed by thematic recommendations for each topic. This is covered from Chapter 3 to Chapter 8, focusing on ethical evaluation and user perspective, data, explainable model development, tools and testing facilities, and AI-based CCAM deployment.
- The final chapters (i.e. Chapter 8 and 9) draws together the recommendations outlined in earlier sections, presenting a consolidated set of integrated guidance for the deployment of AI-driven Connected and Cooperative Automated Mobility (CCAM) systems, followed by conclusion.

2. AI IMPACT AND CHALLENGES

This chapter introduces the growing use of Artificial Intelligence (AI) in the automotive sector. It discusses the role of technologies like Deep Learning and Machine Learning in enabling advanced safety features and addressing perception, prediction and decision making for self-driving vehicles, and standards. Key challenges are highlighted. The chapter identifies the ongoing challenges of using AI with the associated solutions being discussed in the next chapter as project results and outputs.

2.1 Preface

AI applications in the automotive sector, particularly in Automated Driving Systems (ADS), have gained considerable traction in recent years. While AI-based systems, including self-driving cars, are still under development, they leverage technologies like Deep Learning (DL) and Machine Learning (ML). These technologies allow the system to learn from data, in form of various sensor inputs, for example camera images, LiDAR point clouds or RaDAR detections, enabling improved safety and performance. For instance, object recognition systems use multiple layers of processing units to identify objects, such as pedestrians, by interpreting data from vehicle-mounted cameras. This is a critical step toward achieving fully autonomous vehicles, especially at Level 4 and 5 of driving automation, where vehicles are expected to operate without human intervention.

Despite their potential, these technologies still face challenges in decision-making processes, especially in complex and unpredictable traffic situations. These challenges include making decisions based on qualitative traffic rules and adapting to the ever-changing behaviour of road users. As a result, the adoption of AI for these systems continues to be refined to ensure reliability and safety.

New Assessment Test Method (NATM)

Given the increasing complexity of Automated Driving Systems (ADS), traditional vehicle testing methods are inadequate. The New Assessment Test Methodology (NATM) (UNECE 2023) has been introduced by the GRVA (Working Party on Automated/Autonomous and Connected Vehicles) to address these limitations. It provides a flexible and multi-pillar assessment approach for validating ADS, adaptable to various regulatory environments such as EU 2022/1426.

- **Virtual Testing**

Virtual testing serves as the first pillar of NATM, enabling manufacturers to simulate a wide range of driving scenarios that might be difficult or costly to reproduce in the real world. However, it must undergo independent validation to ensure the accuracy of the simulation and its alignment with real-world conditions.

- **Track Testing**

Track testing takes place in controlled environments where real vehicles interact with either simulated or actual obstacles. This testing method helps evaluate system behaviour in critical edge-case scenarios (such as sudden pedestrian crossings, unexpected obstacles, or malfunctioning traffic

signals), providing valuable insights into system safety and performance under controlled yet realistic conditions.

- **Real-World Testing**

Real-world testing ensures that ADS can handle the full spectrum of road and traffic conditions encountered during everyday use. It exposes the system to dynamic variables like human driver behaviour, unexpected obstacles, and environmental challenges, validating the system's ability to perform in real-life situations. Different countries in Europe have distinct road systems, traffic laws, and signage. Real-world testing is important to validate the system's ability to operate effectively within various legal and operational frameworks, ensuring that ADS meets regulatory requirements across borders.

- **Audit Procedures**

Audit procedures involve independent reviews of the entire ADS lifecycle, from design and development to deployment and operation. They ensure that testing methods, safety protocols, and documentation meet the required regulatory standards, providing traceability and accountability for the system's compliance with regulations. Audits generally entail the involvement of independent evaluative bodies that systematically review and confirm that all testing, whether virtual, track-based, or real-world, has been executed in accordance with established standards. This process is essential for ensuring conformity with regulations such as the EU 2022/1426 Act and UNECE standards.

- **In-Service Monitoring**

Once deployed, in-service monitoring continues to track the performance of ADS during real-world operation. This ongoing oversight, along with continuous improvement / deployment, is essential for identifying unforeseen issues and maintaining safety standards over the vehicle's operational life. In-service monitoring is crucial for detecting failures that occur due to unanticipated interactions in real-world conditions, such as software malfunctions, system overloads, or sensor degradation over time. Additionally, in-service monitoring facilitates liability determination by enabling manufacturers, operators, and regulators to track incidents and establish whether the ADS or a human driver was responsible in the event of an accident.

2.2 AI impact on regulations and policies

2.2.1 Regulatory challenges

As AI technologies evolve rapidly, they present new challenges for regulatory bodies in the automotive sector. Traditional certification processes, such as type-approval, were designed for static systems and are often inadequate for AI-driven systems, which can adapt and learn over time. AI models, unlike traditional systems, evolve as they collect more data, which can make it challenging for regulators to ensure these systems remain safe over time without extra oversight. AI models, particularly those using deep learning (DL), pose as "black boxes", making it difficult (even for the developers) to explain how they make decisions. This presents a significant challenge in terms of transparency and accountability, which are crucial for public trust and safety.

AI systems, especially in Automated Driving Systems (ADS), are expected to continuously learn from new data, which complicates long-term safety assurance. Unlike traditional systems that remain constant after deployment, AI-driven systems evolve and may perform differently over time, raising concerns about their ability to comply with established safety and regulatory standards. For example, if an ADS makes an unforeseen decision during an emergency situation, understanding the rationale behind that decision is essential for regulatory compliance and maintaining public trust. AI systems can sometimes push beyond established limits, creating unpredictability in safety-critical functions. Without clear limits and oversight, there is a risk that AI systems might behave unpredictably, especially in novel or untested environments. The EU AI Act (Regulation 2024/1689) addresses some of these concerns by establishing requirements for continuous risk management and transparency in high-risk AI systems throughout their lifecycle. The Act categorizes AI systems based on their risk level, from unacceptable to minimal risk, and imposes stringent obligations on those deemed high-risk. Many automotive AI applications are classified in this high-risk category. These systems must comply with rigorous requirements concerning risk management, data quality, transparency, robustness, and cybersecurity. In the automotive industry, this includes systems integrated into Advanced Driver Assistance Systems (ADAS) and Automated Driving Systems (ADS), which directly impact safety-critical operations.

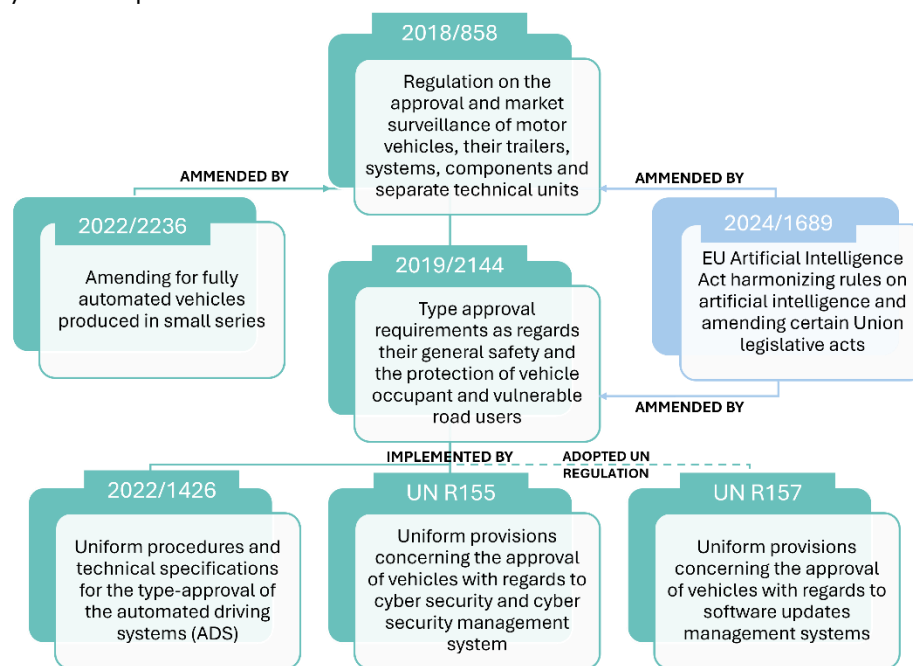


Figure 2: EU Vehicle Regulations Framework with EU AI Act

Regulations such as (EU) 2018/858 (see Figure 2) on the approval of motor vehicles and their trailers, and (EU) 2019/2144 on general vehicle safety, mandate that when adopting delegated acts concerning Artificial Intelligence systems which are defined as a “safety components”, the Regulation (EU) 2024/1689, that aims to close the gaps related with the Artificial intelligence in other regulations, shall be considered. This regulation establishes a direct legal link between AI regulation and sector approval processes, improving the harmonization of AI with automotive legislation. Many other key regulations such as UNECE R155 (cybersecurity), R156 (software updates), R157 (automated lane

keeping), do not yet address the nuances of machine learning or decision-making. R155 and R156 play a foundational role in ensuring that connected and automated vehicles maintain cybersecurity and software reliability, both of which are critical enablers (and vulnerabilities) of AI functionality.

Both regulations focus on establishing general processes for vehicle protection and software management but do not distinguish between traditional systems and those based on AI. In the case of Regulation R155, it sets general requirements for cybersecurity management and threat mitigation, but it does not specifically differentiate AI-based systems from other electronic vehicle systems. The same applies to Regulation R156, which focuses on the processes and systems for managing software updates in vehicles but makes no specific mention of AI technologies or machine learning. For example, manufacturers are required to identify the systems that could be affected by an update, but they are not required to analyse how an update might alter the behaviour of an intelligent system. Nor is there a requirement to specifically assess the risks associated with algorithms or neural networks. This same limitation is observed in other regulations, such as UNECE R157, where there is also no mention of artificial intelligence as an underlying technology, nor is there any requirement for specific validation of algorithms that may learn, adapt, or make complex decisions.

AI testing and deployment also introduces additional challenges, different countries have varying approaches to regulating AI in vehicles, with some favouring self-certification (e.g., the United States) and others, like the EU, relying on type-approval systems. These differences can create inconsistencies in how AI systems are developed and deployed across borders, even inside the European countries. Thus, the EU efforts are to establish a unified regulatory framework that harmonizes standards across member states.

In this context, while UNECE starts new working groups, research projects such as AITHENA are essential for the development of guidance and best practices for future AI regulations. AITHENA bridges technical research with policy development, offering methodologies and tools to enhance transparency, traceability, and auditability of AI models in automotive contexts. This project also addresses a crucial policy tension: while AI enables systems to evolve through data, existing regulations assume static, fully validated software components. AITHENA's proposals such as AI model cards and audit preparation tools aim to help regulators and manufacturers reconcile these gaps.

2.3 AI impact on NATM

The New Assessment Test Methodology (NATM), designed to address the challenges of testing Automated Driving Systems (ADS), needs to evolve in response to the growing complexity of AI-based technologies. Traditional testing procedures, like those used for non-AI systems, are not sufficient for evaluating the dynamic and adaptive nature of AI models. The multi-pillar approach of NATM, which includes virtual, track, and real-world testing, can be adapted to ensure comprehensive evaluation of AI systems.

AI introduces new challenges in real-world testing because of its ability to continuously learn and adapt based on new data. Virtual testing must simulate complex real-world conditions to ensure that AI systems perform reliably in diverse environments. Additionally, in-service monitoring will become more critical as AI systems update themselves over time. Regulatory frameworks must accommodate these updates and ensure that AI systems remain compliant with safety standards even as they evolve.

Transparency, data management, explainability and robustness are to be integrated into the NATM framework to assess AI systems (GRVA 2023). These principles are essential for evaluating how AI makes decisions and ensuring that stakeholders can trust the outcomes of AI-powered systems.

Transparency is crucial for ensuring that AI decisions are both understandable and verifiable. Within the context of Automated Driving Systems (ADS), this entails clearly communicating the vehicle's actions to users, pedestrians, and regulatory bodies. This approach fosters trust among users and regulators and guarantees that AI systems can be audited for fairness and safety. Effective data management is essential for AI, ensuring datasets are diverse to prevent bias and all data is handled securely throughout its lifecycle, from collection to deletion, in compliance with regulations, such as GDPR. Explainability is vital for AI systems, enabling users and regulators to understand and assess AI decisions, which fosters accountability and supports improvements in safety-critical applications like autonomous vehicles. AI systems should be tested for robustness, ensuring that they can operate reliably under unpredictable conditions, such as extreme weather or sensor malfunctions.

2.4 AI impact on safety assurance standards

The integration of AI in ADS presents both opportunities and challenges for safety assurance. While AI can improve safety through advanced perception and decision-making capabilities, its dynamic learning process introduces uncertainty. This uncertainty makes it difficult to predict how AI systems will behave in novel or high-risk situations, raising concerns about system reliability and the potential for unexpected failures. The opaque or "black box" nature of many AI models, such as deep neural networks, makes it difficult to explain their decisions, which is a barrier to public trust and regulatory compliance. AI's continuous learning enhances responsiveness in critical situations but requires rigorous validation before updates are deployed. Evolving standards aim to address safety gaps created by dynamic AI in automotive systems.

Current safety standards, including ISO 26262, which provides a framework for functional safety in electric and electronic vehicle systems, were developed with traditional systems in mind, incorporating structured processes like Hazard Analysis and Risk Assessment (HARA). These standards do not fully cover the specific challenges introduced by AI (see Table 1). For instance, AI systems' adaptive learning and their reliance on large datasets complicate traditional risk assessment methods. As a result, standards like ISO 21448 (SOTIF) and new frameworks like ISO 8800 are emerging to address AI-specific risks, including system transparency, robustness, and the need for continuous validation as AI models evolve. ISO 21448 addresses safety concerns in advanced driver

assistance systems (ADAS) and autonomous driving systems that go beyond traditional hardware failures covered by ISO 26262.

Furthermore, standards such as ISO/PAS 8800:2024 address the risks related to safety at the vehicle level caused by output insufficiencies, systematic errors, and random hardware errors of AI elements within the vehicle. This includes interactions with AI elements external to the vehicle that may directly or indirectly affect vehicle safety. While it encompasses principles from established and well-understood classes of machine learning (ML), it does not delve into specific AI methods such as deep neural networks. Moreover, it lacks specific guidelines for software tools that incorporate AI methods. The scope of the standard applies to electrical and/or electronic (E/E) systems in series production road vehicles that incorporate AI technologies, particularly machine learning (ML). It excludes mopeds and special vehicles designed for drivers with disabilities.

In parallel, process-oriented frameworks like ASPICE 4.0 (derived from ISO/IEC 15504), which assess the maturity of automotive software development, are also incorporating dedicated guidelines for AI projects. ASPICE 4.0 includes Machine Learning Engineering processes, but full traceability to ISO 26262 validation goals is still evolving. ISO/TS 5083:2025 offers comprehensive guidance for ensuring and demonstrating the safety of automated driving systems (ADS) integrated into road vehicles. This approach is grounded in globally recognised safety principles and top-level safety objectives. It encompasses safety by design, verification and validation processes, and post-deployment activities for level 3 and level 4 ADS features, as defined in ISO/SAE PAS 22736. The standard is specifically intended for road vehicles, including buses and trucks, while it excluding motorcycles and mopeds.

Table 1: Comparison of Key Automotive Safety Standards in Addressing AI-Related Risks and Limitations

STANDARD	WHAT IT COVERS REGARDING AI	WHAT IT DOESN'T COVER
ISO 26262	Provides a structured approach to risk analysis (HARA) for traditional systems.	Does not adequately address AI-specific risks such as model opacity, robustness to novel data, and dynamic adaptability.
ISO 21448 (SOTIF)	Considers operational safety even in the absence of technical faults. Relevant for validating safe behaviour of AI in unpredictable scenarios.	Validating AI models trained with limited data is a challenge; the standard does not offer specific methods to ensure safe generalization.
ISO 8800:2024	Aims to address existing gaps by offering a dedicated framework for safety-critical AI systems.	It does not cover AI methods like DNN or provide guidelines for software tools using AI.
ASPICE 4.0	Includes guidelines for engineering systems with Machine Learning, including processes adapted to the ML lifecycle.	Needs stronger integration with functional safety standards and specific criteria for AI validation and certification.
ISO/TS 5083	It focuses on ML, as it is one of the core components of ADS and	Requires detailed guidance on model updates and data drift

	provides definitions of safety-related properties of the AI system.	
--	---	--

2.5 Challenges

Table 2 outlines the principal challenges faced in building trustworthy, interpretable AI systems and presents approaches (and solutions, where applicable) to address these issues, drawing on references from the AITHENA project. The information herein is intended to guide stakeholders in understanding the current landscape and ongoing efforts in ensuring safe and reliable AI deployment.

Table 2: Challenges and approaches to build trustworthy and interpretable AI systems

CHALLENGES	APPROACHES
Building trust through interpretable systems	Model monitoring point 2.1.4.2. (AITHENA D3.1) UC3 and UC4 Point 4 (AITHENA D5.1) KPI_10, KPI_11, KPI_12, KPI_13, KPI_14, KPI_15, KPI_16, KPI_17, KPI_18 (AITHENA D5.1)
Need for continuous monitoring of safety-critical functions	
AI models continuously evolve as they gather more data	
Providing clear reasons for AI decisions to users and regulators	
Ensuring AI decisions are understandable and verifiable to all stakeholders	
Auditing decision pathways for safety assurance	
Supporting accountability when systems make errors	
Addressing the "black box" nature of deep neural networks	Training point 2.1.2 (AITHENA D3.1) Explainability methodology point 3.3 (AITHENA D1.1) UC1 point 4 (AITHENA D5.1) KPI_20, KPI_21, KPI_22, KPI_23, KPI_24 (AITHENA D5.1)
Building trust through interpretable systems	
Clarifying what went wrong and how to correct failures	
Updates based on new data can affect safety-critical functions without prior verification	
Enabling continuous improvement through understandable failures	
Assessing trustworthiness despite limited transparency	Trustworthy AI-enabled CCAM Applications: Lawful, ethical, and robust point 1.2 (AITHENA D1.1) Societal Fairness: Safety point 2.2.2 (AITHENA D1.1) UC3 point 4 (AITHENA D5.1)
Additional oversight beyond initial certification	
Risk of unpredictable behaviour in novel or untested environments	
Ensuring datasets are diverse and representative to prevent biased decision-making	

Establishing appropriate limits for autonomous decision-making	KPI_10, KPI_11, KPI_12, KPI_13, KPI_16 (AITHENA D5.1)
AI systems may push beyond traditional safety limits unlike conventional ADAS	
Difficulty predicting responses to novel situations	
Maintaining compliance with data regulations like GDPR	<p>Lawfulness point 1.2.1 (AITHENA D1.1)</p> <p>UC2.1 and UC2,2 (AITHENA D5.1)</p> <p>KPI_01, KPI_02, KPI_03, KPI_04, KPI_05, KPI_06, KPI_07, KPI_08, KPI_09, KPI_19</p>
Cross-border inconsistencies in development and deployment standards	
Creating flexible yet robust frameworks for rapidly evolving technology	
Challenges creating globally recognized safety standards	
EU type-approval approach differs from other regulatory frameworks	
Operating reliably across varying environments	

3. ETHICAL EVALUATION AND USER PERSPECTIVE

This chapter addresses the challenge of assessing trustworthy AI in CCAM solutions via development of a human-centric methodology in AITHENA highlighting the checklist as a major output of the project along with key insights and the user mobility needs, expectations and concerns regarding AVs.

3.1 Ethical evaluation framework for AI in CCAM

Deliverable D1.1 of AITHENA reviews existing ethical frameworks and adapts them to address the specific challenges of CCAM applications. It proposes new frameworks as necessary to address gaps in current methodologies. It includes [methodological checklist](#) that make concepts such as fairness, accountability, privacy, and transparency practical and actionable for developers and regulators. This covers detailed guidelines for evaluating the social impact of AI technologies and ensuring they contribute positively to societal well-being.

3.1.1 Key insights and lessons learned

Evaluating fairness:

- Fairness in CCAM must extend beyond algorithmic bias to encompass societal equity, transport accessibility, and universal design.
- Meanwhile, bias can arise at multiple levels—data collection, algorithm design, and user interaction—necessitating comprehensive mitigation strategies. Most research centres on race and gender, overlooking less obvious or emerging biases. Social structures inherently contain bias, and AI systems must account for real-world complexities.
- Fairness is subjective and varies by context, making universal solutions difficult. Evaluating fairness should therefore be done through technical as well as societal perspectives, exploring different sub-components such as equality, safety, accessibility, procedural fairness, bias, and non-discrimination.
- Fair ML tools have emerged in the past decade, supporting fairness exploration and evaluation. These include:

FairLearn: Offers tutorials and code examples.

Microsoft AI Fairness Checklist: Covers fairness across the product lifecycle.

FairSight: Visual dashboard for fairness evaluation.

AI Fairness 360: Open-source toolkit with bias mitigation algorithms.

Evaluating transparency:

- The level of transparency required depends on the criticality of the task. For safety-critical applications (e.g., autonomous driving), high transparency is essential.
- Explainable AI (XAI) methods help users and regulators understand AI decisions. Local explanations (e.g., why a specific decision was made) and global explanations (e.g., how the model generally behaves) are both important.

- Different dimensions of transparency should be taken into consideration around every ML model. These include:
 - Input transparency:* Understanding which inputs influenced a decision.
 - Model transparency:* Understanding how the model processes inputs (via model cards).
 - Dataset transparency:* Knowing how data was collected, processed, and its limitations (via data cards).
 - Communication transparency:* Tailoring explanations for different audiences (developers, users, regulators).
- There are several challenges in achieving transparency. While black-box models (e.g., deep learning) are powerful, they are hard to interpret. Post-hoc explanations can help but may not fully reveal model logic. In addition, data quality and documentation are often insufficient for full transparency.

Evaluating accountability:

- Accountability involves assigning ownership and responsibility for AI systems and ensuring that actors can explain and justify their actions. It is relevant includes responsibility, accountability, consultation, and information-sharing roles (RACI matrix).
- The accountability of AI-based products, services, or systems should be considered based on specific roles in the building, deployment, and maintenance process. Key actors to consider for CCAM applications include developers, integrators, operators, maintainers, and regulators.
- Data traceability is critical for accountability tracking data from source to use, including transformations and access logs. Tools like DVC, DagsHub, MLFlow, Weights and Biases, and Neptune offer tracking of data, experiments and models for data version control.
- Regulations exist for functional safety testing, but testing AI decision-making systems presents challenges because these systems are not deterministic.

Evaluating privacy:

- AI systems in CCAM often process personal and sensitive data (e.g., location, biometrics, driving behaviour). The General Data Protection Regulation (GDPR) is the primary legal framework governing data protection in the EU.
- It is imperative that developers incorporate privacy protections from the initial stages of system design. This includes principles such as data minimisation, purpose limitation, and ensuring user control over personal data. To uphold privacy standards and regulatory compliance, notably with regulations such as the GDPR, it is crucial to employ techniques such as data anonymisation, federated learning, or pseudonymisation. Federated learning permits model training without centralising data, thereby mitigating privacy risks. Moreover, it is crucial to identify scenarios where pseudonymisation is more appropriate than anonymisation, and vice versa.
- Explicit, informed consent is required for processing personal data. Users must be able to access, correct, or delete their data.

- CCAM systems involve complex data flows between vehicles, infrastructure, and cloud services. Data sharing must be secure, justified, and transparent, especially when involving third parties.
- GDPR may not fully address the nuances of AI and CCAM (e.g., data from passengers or bystanders). The lack of differentiation between private and shared vehicles further complicates the applicability of privacy for different vehicles and uses.
- Edge devices and infrastructure also collect data but are often overlooked in privacy assessments.

3.1.2 Recommendations

- **Checklist timeline:** The checklist via D1.1 provides evaluation framework for trustworthy elements regarding AI in CCAM solutions. This evaluation will vary depending upon different phases of a solution which is being developed i.e. from development to operations. This should be taken into consideration.
- **High-value datasets:** Developers, auditors, and policymakers should work to identify high-value datasets that should be protected as public resources as well as datasets that would improve CCAM system development and safety if shared across stakeholders.
- **Mandate transparency standards:** Transparency standards should be mandated requiring transparency documentation for all AI systems in CCAM, including feature importance rankings, dataset provenance and explanation methods used.
- **Adopt explainability guidelines:** A comprehensive list should be compiled of the Explainable AI (XAI) methods employed to derive insights from a black-box model. Each method should be itemized individually, with its corresponding insights highlighted and the level of explanation (local or global) specified. This list should also encompass any machine learning models that inherently provide explainability (white-box models) along with their respective insights. Encourage the use of XAI toolkits and checklists (e.g., ALTAI, IEEE P7001).
- **Tailor communication to stakeholders:** Develop user-friendly explanations for non-technical audiences. Ensure regulators receive full technical documentation and audit trails.
- **Liability and certification:** The relationship between certification processes (to meet safety standards) and liability requires refinement to prevent situations where certain self-driving functions are safety-certified but not legal in some EU Member States, or where they may be legal but not certified due to lack of legal definition or grey areas. Approach should be to develop and mandate a European certification scheme (protocol) for AI-based CCAM, in alignment with the EU AI Act and GDPR, involving multi-stakeholder input.
- **Data traceability:** In the context of data used for CCAM, developers should use a tool that registers all versions of datasets that are used for the training of the different models used in the vehicles. This includes data lineage tracking from collection to deployment.
- **Support testing and simulation:** Standardised test protocols should be developed for AI behaviour under uncertainty with investments in scenario-based testing platforms and simulation environments for AVs.

- **Cross-border regulatory alignment:** International cooperation on AI and CCAM standards should be facilitated by encouraging mutual recognition of certifications and testing results across EU member states (and global partners).
- **Clarify consent mechanisms for privacy:** Develop standardised consent protocols for shared and public transport scenarios. Ensure passengers and bystanders are informed and protected, not just drivers.
- **Promote privacy-preserving technologies:** Encourage adoption of federated learning with proper encryption in a way that doesn't impact the quality of the results negatively.
- **Expand GDPR to address AI-specific risks:** GDPR interpretations for AI-driven mobility systems should be provided. This should address edge cases like data capture of non-consenting individuals in public spaces.
- **Data protection officers for CCAM:** Mandate appointment of DPOs for all CCAM operators, defining the data limits for processing.

3.2 User needs, expectations and concerns

The developed CCAM solutions concerning AI will affect a range of stakeholders, including autonomous vehicle (AV) users, non-AV users such as pedestrians and cyclists, CCAM solution developers, and decision-makers. It is essential to understand key requirements from these user groups, focusing on their mobility needs, expectations, and concerns related to CCAM implementation. This is outlined in AITHENA's [Deliverable D1.2](#).

3.2.1 Key insights and lessons learned

User needs and expectations

- Considering the potential use of AVs in various scenarios, the concept of automated public transport vehicles with fixed routes, stops, and high frequency shows significant public interest. Additionally, the use of an automated shuttle service (similar to on-demand shuttle) also attracted considerable interest. Using AVs for commuting as a passenger was preferred over owning or having access to shared self-driving vehicle services.
- Low level income groups (i.e. <10,000€) had least interest in owning a personal AV. In addition, owning or having current access to car showed high interest in owning a self-driving vehicle.
- Trust in AVs is notably lower in rural and moderate in urban and suburban areas compared to high level of trust in inter-city areas (e.g., highways), reflecting concerns about AVs' reliability in urban environments.
- There are a significant interest and expectation in AVs improving travel time and safety, with less traffic congestions, and the AV system being environmentally sustainable.
- There is a high expectation for transparency regarding AV decisions, particularly in how the vehicle perceives its surroundings, such as pedestrians and obstacles. Strong emphasis is observed on the importance of knowing what information the AV's sensors detect and how confident the vehicle is in its decision-making.

Concerns

- Concerns regarding the safety and reliability of self-driving vehicles in extreme conditions, user privacy protection, and cybersecurity threat were identified.
- The main factors discouraging self-driving vehicle use are equipment or system failure, cyberattacks, and legal liability in case of a crash.

3.2.2 Recommendations

- **Support AV integration into public transport systems:** Governments should incentivize local authorities to integrate AVs into public transport fleets. This could include funding for infrastructure upgrades such as dedicated AV lanes, charging stations, and integration with existing public transport networks. Policies should also focus on ensuring that AV-based public transport is affordable and accessible to all users, particularly in underserved areas.
- **Building trust and enhanced transparency:** Policies should prioritise the development of systems that provide clear, understandable information to users about AV decision-making processes, sensor capabilities, and data usage. Development of AI models should be done with clear transparency, accountability, and explainability.
- **Safety assurance in diverse conditions:** AV manufacturers should conduct real-world testing across a range of environments, from urban streets to rural roads, in various weather conditions, and in mixed-traffic scenarios. Testing should focus on AVs' ability to handle complex, unpredictable environments.
- **Liability guidelines:** A clear legal framework for AVs should outline liability in accidents or malfunctions, defining roles of manufacturers, developers, and consumers, and include a mechanism for insurance and compensation (in case of accidents).

4. DATA - Life Cycle Management and Generation

This chapter explores the challenges surrounding AV reliability, user privacy, and cybersecurity, while highlighting the necessity for robust data governance and provenance mechanisms. By examining policy recommendations, legal frameworks, and advanced tools for data and model documentation, the chapter aims to provide a perspective on fostering trust in AV technologies.

4.1 ML DevOps Data governance and provenance mechanisms

4.1.1 Key insights and lessons learned

Data governance and provenance:

- Data governance and provenance ensure transparency, accountability, and traceability in machine learning (ML) models, which is critical for high-stakes domains like CCAM. These mechanisms help mitigate risks such as biased AI decision-making, privacy breaches, and unsafe data practices.
- The framework supports the EU's AI Act by embedding compliance measures and promoting ethical AI practices, such as privacy preservation and bias reduction.
- Tools like ClearML and FiftyOne facilitate efficient data versioning, monitoring, and visualisation, while Data Cards (introduced in the project) ensure consistent documentation of datasets, enhance data traceability, crucial for compliance, reproducibility, and trustworthiness of AI models in CCAM applications.

Data cards, Model Cards, and MLOps Cards:

- AITHENA's Data Card details dataset's origins, preprocessing, limitations, and ethical considerations for CCAM. They improve transparency and help assess data quality, provenance, and biases.
- Model Cards outline AI models' training, performance metrics, and caveats, aiding developers and regulators in understanding model behaviour in real-world scenarios.
- MLOps Cards offer insights into AI models' deployment and operation, including maintenance, retraining cycles, and performance monitoring.

Data quality and version control:

- Ensuring high-quality data and robust data governance practices, including continuous monitoring of data drift and model performance, is a central principle of data operations (DataOps). Techniques such as statistical tests for feature distribution changes and model performance tracking help ensure that AI models remain reliable over time, even as operational data evolves.

Tools and methodologies:

- Tools like ClearML for version control, FiftyOne for data visualisation, and Foxglove Studio for multi-sensor data exploration assist for an efficient data management and model development in MLOps. They enable seamless collaboration, continuous integration, and AI model deployment in CCAM systems.

4.1.2 Recommendations

- **Enhance regulatory frameworks:** Authorities should adopt and integrate Data Cards and Model Cards within the regulatory framework to ensure AI systems in CCAM meet transparency and accountability standards. This will help stakeholders- especially regulators- understand the data and models driving AV systems, enabling informed decision-making about safety and compliance.
- **Establish data governance best practices:** Stakeholders should implement data governance frameworks that focus on continuous data versioning, quality assurance, and drift detection. This includes leveraging tools like ClearML for dataset versioning and FiftyOne for dataset exploration and curation, ensuring data integrity and alignment with ethical and regulatory guidelines. Authorities should prioritise the establishment of continuous monitoring systems and adaptation through investment for AI models.
- **Encourage collaboration across stakeholders:** Promote cross-disciplinary collaboration among AI developers, data scientists, regulators, and ethicists through standardised tools like Data Cards and Model Cards.

4.2 Privacy Preserving ML

4.2.1 Key insights and lessons learned

- The difference between anonymization vs. pseudonymization is vital for privacy protection. Anonymization, being irreversible, ensures data cannot be re-identified, while pseudonymization is reversible with access to the necessary "key". Understanding the implications of each is critical for ensuring that privacy is maintained while still enabling AI models to function.
- Techniques like blurring, deep fake, and generative substitution are emerging as ways to anonymize sensitive data (e.g., faces, license plates) in the context of CCAM, though these methods may influence the accuracy of AI models.
- Federated Learning (FL) offers a potential solution to ensure privacy in data processing by allowing AI models to be trained across multiple devices (e.g., vehicles) without transferring sensitive data. However, the decentralized nature introduces risks such as data poisoning and model manipulation, which must be addressed through robust security protocols.

4.2.2 Recommendations

- Authorities should mandate the integration of privacy-preserving techniques such as anonymization and pseudonymization in AI systems, particularly in CCAM solutions, ensuring that AI developers adhere to them in their design processes.
- Policymakers should establish clearer frameworks for obtaining informed consent in CCAM systems, for individuals inside and outside the vehicle (e.g. pedestrians). Development of dynamic consent management tools and notifications that inform people about data collection in real-time should be encouraged.
- Governments should incentivize the adoption and research of Federated Learning (FL) for CCAM, ensuring it aligns with privacy and security requirements
- Authorities should require CCAM AI developers to provide clear data policies and set up user-friendly interfaces that allow individuals to understand and manage their privacy settings (e.g., information about what data is collected, how it's used, and for how long it's retained).

5. XAI - Explainable Continuous Model Development

The chapter provides key insights, an overview of frameworks and tools that enhance transparency and trust in machine learning systems, and recommendations for policymakers and authorities. Readers will gain a comprehensive understanding of the challenges and solutions shaping responsible AI development in CCAM.

5.1 ML Development and life cycle management framework

5.1.1 Key insights and lessons learned

- Comprehensive ML Lifecycle Structure: Machine learning (ML) is the process by which AI algorithms are developed. The ML framework (outlined in Deliverable D3.1), involves four stages in its process: Data, Training, Testing, and Deployment, which makes up the lifecycle of a ML model. D3.1 focuses on the methodology and tools for the development and life-cycle assessment of individual ML algorithms. It includes a list of main tools for data, training, testing, deployment, and general development for ML life cycle management.
- Model Cards for transparency and trustworthiness: The introduction of model cards standardises documentation for AI models, providing insights into model architecture, training/evaluation data, intended uses, ethical considerations, and potential biases, advancing explainable AI (XAI) principles critical for CCAM safety and trust. The AITHENA model card is designed to address issues such as fairness, with built-in bias mitigation techniques, and it allows document includes these issues for privacy protection and accountability. These measures ensure that AI decision-making is transparent, providing a clearer explanation of why a system (such as a self-driving car) makes one decision over another. The aim is to demystify the 'black box' problem in AI, where decision-making processes lack transparency or logical explanation. The AITHENA Model Card includes sections such as:
 - (1) Introduction (which includes contact details of the author, affiliation, etc, and a summary of the model, relevant information as well as license),
 - (2) Model details (provides more detailed information about the model, including its architecture, task, inputs and outputs. Information about the data used to train and test the model is also requested with sub-sections on model architecture, training data, and evaluation data)
 - (3) Intended use (describes the intended use of the model. Mentioning possible scenarios of implementation, the description of the users expected to use the model and the ones the model will be analysed. The subsections include use cases and users, and limitations)
 - (4) Evaluation and performance (providing information on metrics and subjective observations about the performance of the model: accuracy, confusion matrix, etc.)

- (5) Ethical consideration (includes information of how the model is including the aspects of fairness, privacy and accountability. It can include the report of fairness metrics or bias; documentation of how the data is processed)
- (6) Usage and risks (includes detailed instructions and recommendations for using the model. It highlights the risks and limitation when using the model, this can include ethical or technical issues)
- Use of synthetic data for robustness assessment: The framework (in D3.1) demonstrates practical application of synthetic data in robustness testing. Synthetic data is artificially generated and is often used in scenarios where collecting real-world data is challenging for reasons such as privacy concerns, rarity of events, or cost.

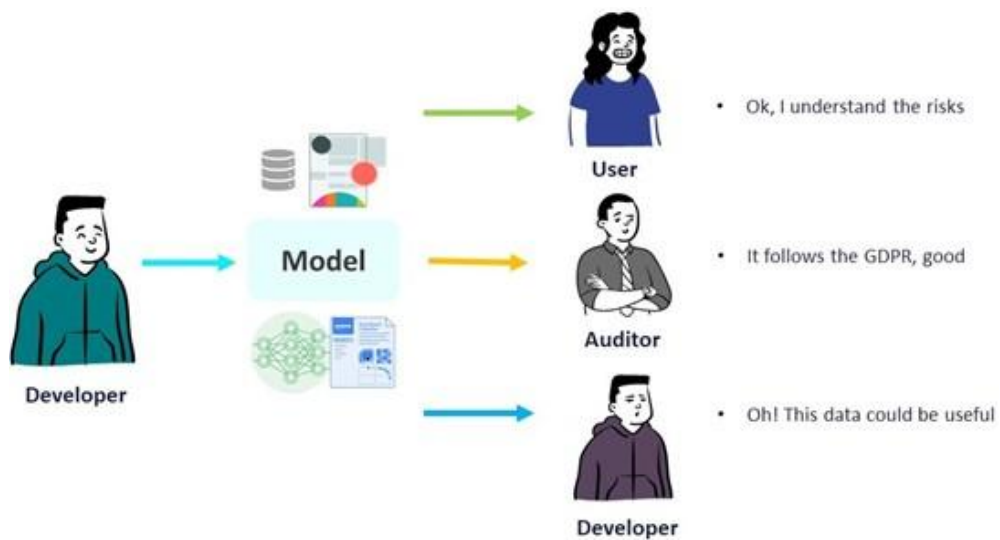


Figure 3: Image showing how developers of AI systems can communicate relevant information about the system to different users via Model Cards (Source: [AITHENA 2023](#))

5.1.2 Recommendations

- **Implement ML Lifecycle Transparency via Model Cards:** Regulatory bodies should establish requirements for standardised model cards to accompany every deployed AI system in CCAM.
- **Prioritize Ethical AI:** Use AITHENA framework as a compliance benchmark using checklists and reference to tools from AITHENA D1.1 and D3.1.
- **Encourage Use of Explainable and Hybrid AI Models in Safety-Critical Applications:** Require explainability reporting for AI models in high-stakes CCAM applications. This improves interpretability and reliability in critical CCAM tasks (e.g., pedestrian detection, trajectory prediction), reducing black-box risks.

5.2 AI algorithm development

5.2.1 Key insights and lessons learned

Data and information fusion to reduce conflicting perception

- Importance of multi-sensor fusion: Autonomous vehicles (AVs) depend on multiple sensors (LiDAR, Radar, Camera) to perceive their environment. Each sensor has strengths and weaknesses (e.g., cameras are sensitive to lighting, LiDAR struggles in adverse weather), and by combining data from multiple sensors, AVs can create a more reliable and accurate understanding of the surroundings. Effective fusion methods, such as the LiDAR-Camera Fusion or LiDAR-Radar-Camera Fusion, help address sensor limitations (e.g., limited visibility conditions) and minimise conflicting perceptions, particularly in challenging situations such as adverse weather or reduced visibility.
- Explainable perception and multi-modal transformer models: One challenge in autonomous vehicle (AV) development is explainability. The capability to clarify how and why a system perceives objects in specific ways is relevant for operator understanding and regulatory approval. Implementing Explainable Artificial Intelligence (XAI) techniques in perception systems can increase their transparency. In AITHENA, through the perception modules (including camera based 3-D object detectors), and multi-modal transformer model (fusing camera and LiDAR), the method generates saliency maps that visualizes which inputs are (through highlighted pixels, heatmaps) important for prediction and decision-making.
- Uncertainty quantification: By quantifying the uncertainty of perception models (using techniques like OCCUQ for 3D occupancy prediction), autonomous systems can provide more reliable decision-making in uncertain situations, allowing operators to understand when a system may be less confident about its decisions (or when the system leaves it's ODD).
- Robustness: Perception systems are required to operate effectively under adverse conditions, such as poor weather or sensor misalignment.
- Local Dynamic Maps (LDMs): The integration of perception system outputs in real-time from different data sources (e.g., map information, 3D object detections, V2X messages, traffic signs, and traffic light signal phase) into LDMs enables AVs to dynamically update their understanding of the environment, facilitating more accurate navigation and decision-making.

Edge cases and use of data

- Edge cases, such as rare objects, unusual driving conditions, or unexpected sensor failures, can be challenging for autonomous driving systems. Addressing these situations requires focused training and modelling to enable proper system response. Training and testing with both real and synthetic data are used so that AVs can operate effectively in irregular or infrequent scenarios.
- Considering situation awareness on enhanced object detection, extreme weather conditions and high vehicle speeds (above 80kmph) can negatively impact detection reliability, resulting in poor-quality outputs. Onboard cameras may unintentionally record identifiable public data, so proper anonymisation is needed to address privacy concerns.
- In the virtual world simulation approach, synthetic data is generated to represent various critical events that are challenging to capture in real-world scenarios. This method involves

the use of Radar and LiDAR sensor models, which can be complex and may require specific technical knowledge to understand.

Explainable and robust decision making

- A hybrid approach in AI motion planning improves both trajectory quality and system stability, which is critical for real-world application.
- The ability of AI systems to assess their own competence in complex or unfamiliar scenarios is vital. AITHENA' s work includes developing competence estimation and risk assessment models using knowledge graphs to enhance decision-making in dynamic environments.
- AI models for traffic management (e.g., VISSIM/SUMO simulation) demonstrate the importance of understanding network-level impacts of AVs on urban traffic.

5.2.2 Recommendations

- **Promote multi-sensor fusion standards:** Governments and regulatory bodies should establish standards for sensor fusion algorithms in AVs. These standards could provide guidance on the integration of data from different sensors (e.g., LiDAR, radar, camera) to ensure the reliability and consistency of AV perception systems.
- **Regulate explainability:** Regulatory frameworks should make sure that AV perception models are explainable by including XAI methods to provide clarity in the reasons for system's decision-making. Authorities should define minimum explainability requirements for predictive AI models (e.g., visual evidence for decision-making).
- **Edge-case training and validation:** Investments in the collection and use of synthetic data (along with real data) should be encouraged to simulate edge cases in training autonomous systems.
- **Promote Hybrid AI approaches:** Support the development of hybrid AI models through funding initiatives and facilitating partnerships between research institutions, private companies, and government agencies. Consider implementing incentives to encourage organisations to adopt these systems in both simulation environments and real-world testing.
- **Simulate AI integration:** At a network scale, AI integration should be simulated before approving large-scale AV deployment. AI systems must adapt to mixed traffic environments, with algorithms capable of adjusting driving behaviour based on trust levels and traffic conditions.

6. TOOLS AND TESTING FACILITIES

Innovations in autonomous vehicle technology are transforming the future of mobility. This chapter explores essential strategies, regulatory priorities, and the pivotal role of rigorous testing to support the advancement and safe integration of connected and automated vehicles.

6.1 Physical infrastructure and vehicle setup

6.1.1 Key insights and lessons learned

- Demonstrator vehicles like Kia Niro HEV were selected as the primary development platform (“CAVRide”) due to prior expertise and compatibility with CAN bus protocols. Kia EV6 was primarily used by Siemens-BE due to its ability to draw power directly from the vehicle's battery, providing energy supply to the perception sensors installed on the roof rack as well as to the data acquisition system. Ford Fusion, utilized as VIF's demonstrator test vehicle, was selected due to its adaptability for hardware modifications and its capability to support scenario-awareness functionalities, both essential for testing advanced perception systems. A Volkswagen Multivan eHybrid was selected as the basis for ika's test vehicle, due to its Drive-by-Wire interface allowing software-based vehicle guidance with closed-loop control (AITHENA D4.2).
- The CAVRide was transformed into a Connected and Automated Vehicle (CAV) demonstrator, integrating modular subsystems for perception, control, and V2X communication.
- A robust sensor suite was installed, including LiDAR (Ouster OS128), radar, industrial-grade cameras, Mobileye EyeQ8, and a high-precision GNSS/INS system for accurate localisation and validation.
- Data logging and synchronization were managed through ROS₁/ROS₂ and converted into MCAP format for interoperability with project partners.
- The setup supported real-time testing with displays for engineers, high-speed connectivity (5G/Wi-Fi), and power stability up to 1000 W without external sources.

6.1.2 Recommendations

- Continue refining modularity in hardware and software to facilitate easier integration of new sensors or algorithms.
- Explore next-generation LiDAR and radar with enhanced resolution and range for more complex environments.
- Strengthen redundancy in localization (combining GNSS/INS with vision-based SLAM) to ensure robust ground-truth validation in GPS-denied conditions.
- Develop standardized sensor calibration protocols to improve interoperability between project partners.

- Expand the data management pipeline (e.g., automated cloud upload) to accelerate collaborative research and reduce manual processing.

6.2 Virtual tools and testing

6.2.1 Key insights and lessons learned

- A XiL (X-in-the-Loop) testing environment has been established, incorporating Driver-in-the-Loop (DiL) and Vehicle-in-the-Loop (ViL) configurations, enabling the integration of real human operators, sensors, and Electronic Control Units (ECUs) with high-fidelity digital twin simulations. This setup allows for realistic, real-time interaction between virtual and physical systems, providing a versatile platform for evaluating vehicle and AI-based behaviour and system responses under a wide range of scenarios.
- The Dynamic Vehicle-in-the-Loop (DynViL) system integrated a real vehicle with a virtual environment (using CARLA simulator) to safely and cost-effectively test Automated Driving Functions (ADFs).
- This hybrid approach allowed capturing real vehicle dynamics while simulating complex, high-risk, or costly scenarios.
- The setup minimized on-board computational loads by offloading heavy simulations to external systems via ROS2 communication.
- A simplified ADF was implemented with PID-based control for steering, braking, and acceleration, validated through a benchmark scenario.
- Early tests confirmed the viability of mapping physical vehicle dynamics into virtual simulations, bridging the gap between digital and physical testbeds.

6.2.2 Recommendations

- Support XiL test environment, as it represents a significant milestone in the development of robust and trustworthy AI-based CCAM systems and contributes meaningfully to the advancement and eventual certification of safe, intelligent mobility technologies.
- Enhance the fidelity of virtual scenarios by incorporating 3D models of test tracks and vehicles for closer alignment with real-world conditions.
- Expand the computational setup to allow distributed simulation, addressing delays introduced by wireless communication in complex scenarios.
- Integrate more advanced ADF prototypes (e.g., with predictive collision avoidance or multi-agent planning) to test the system's scalability.
- Develop a library of standardized, reusable virtual test cases to benchmark algorithms consistently across iterations.
- Investigate hybrid sensor emulation (mixing real and virtual sensor feeds) to validate algorithms under diverse conditions before field deployment.

7. DEPLOY & TEST - AI-based CCAM Deployment

This chapter presents a comprehensive overview of the main findings, lessons learned, and strategic recommendations derived from five distinct use cases (with the second use case split in two) explored within the AITHENA project. The content is structured to summarise the outcomes of each use case, highlighting both technical and methodological advances, as well as challenges encountered during development and deployment.

7.1. UC1: Trustworthy Perception Systems for CCAM

7.1.1 Key Insights and lessons learned

- **Trustworthiness requires a holistic, multi-level strategy.** A robust and trustworthy perception system cannot be achieved by focusing on the AI model alone. The project demonstrates that a comprehensive approach combining model-driven (e.g., explainable layers), data-driven (e.g., Data Cards), and sensor-driven (e.g., reliable fusion) strategies is essential to build trust across the entire system stack.
- **Performance on standard benchmarks is an insufficient indicator** of real-world safety. Our evaluations demonstrated that current high-performing perception models exhibit significant performance degradation under adverse conditions like inclement weather or sensor misalignment. This necessitates a systematic evaluation of model robustness beyond standard accuracy metrics to ensure safe operation.
- **Mechanisms to enhance AI transparency** are computationally feasible for real-time deployment. It is possible to generate explainability outputs (e.g., saliency maps showing model focus) and quantify model uncertainty at runtime. This provides critical information for system monitoring and enables downstream functions, like motion planning, to adapt to situations of low model confidence.
- **A trade-off exists between model performance and trustworthiness.** The integration of mechanisms such as uncertainty quantification introduces small computational overhead. This balance must be carefully engineered and validated to meet both safety and real-time performance requirements.
- **Comprehensive and standardized documentation** is a cornerstone of AI trustworthiness. Formal documentation, such as Model Cards and Data Cards, provides essential transparency regarding a model's performance, its limitations, and the characteristics of its training data. This structured approach is fundamental for traceability, accountability, and the certification process.
- **The outputs from transparency mechanisms must be made interpretable to be effective.** An XAI (Explainable AI) Interface is required to translate raw technical data, like model attention or uncertainty scores, into an accessible format. Such an interface is a critical tool for engineers, testers, and auditors to properly understand and validate the AI system's behaviour.

7.1.2 Recommendations

- **Establish standardized testing protocols for AI / CCAM robustness:** In addition to standard benchmarks, certification should require evidence of system performance under a defined set of adverse conditions, including sensor failures, misalignments, and challenging environmental scenarios.
- **Require logging of AI / XAI's output:** Safety-critical perception systems should be required to output and log uncertainty and explainability data. This information is invaluable for post-incident analysis and continuous system validation, serving a similar function to an event data recorder.
- **Promote the adoption of standardized AI documentation:** The use of Model Cards and Data Cards should be a required practice across the automotive supply chain to ensure transparent communication of an AI component's capabilities, limitations, and data dependencies, thereby supporting integration and regulatory oversight.
- **Support the development of open standards for XAI data:** Standardized formats and APIs for exchanging explainability and uncertainty information would foster an ecosystem of interoperable validation and monitoring tools. This would allow regulators and third parties to assess systems from different suppliers using a consistent methodology.
- **Promote standards for human-system interaction in failure scenarios:** Policies should encourage the development of standardized ways for an automated system to communicate its operational state and limitations to the user, particularly when it encounters an unresolvable issue and needs to transition to a safe mode.

7.2. UC2: AI-extended situational awareness/understanding

7.2.1. UC2.1 Collision Prediction with Hybrid AI data fusion models

Key insights and lessons learned

The use of Vision Language Models (VLMs) combined with a formal ontology offers several advantages that can leverage safe and explainable autonomous systems, moving beyond simple object detection to a more holistic, knowledge-driven scene understanding.

- **Handling Out-of-Distribution (OOD) Situations:** OOD is an issue in the real world, where a vehicle might encounter an unusual object (e.g., an unexpected animal on the road, a unique type of road) that could lead to a catastrophic failure if not properly identified. VLMs are trained on a massive and diverse corpus of images. This gives them an expansive, generalized "world knowledge" that goes far beyond the limited scope of a pre-defined training dataset for autonomous vehicles. This inherent capability allows them to identify and describe novel or unforeseen objects and situations.

- **Deeper analysis and explainability:** A reasoning log can be implemented. This shows a “thought process” that, while it cannot be considered as a sufficient technical explanation of the model’s decision, can help build logical paths from the image to the output of the model. It allows a human operator or engineers to instantly see that the system detected "a pedestrian is crossing a road", and not just that "a pedestrian is present", and it matches the assigned risk or general understanding of the scene (e.g., “medium risk: a pedestrian is crossing while the car is moving forward”).
- **Agentic Architecture:** Although this was not finally implemented, initial exploration showed that adopting an agentic architecture (i.e., system containing/coordinating AI agents) adds a layer of control and validation. Relevant for safety-critical applications like AVs. This allows for a "chain-of-thought" process or the decomposition of complex tasks into smaller manageable sub-tasks that can be better supervised or controlled. Also, an initial validation of the input and final validation of the output can be done to ensure that the system has the relevant and correct data to work, and that the system has produced a proper and correct response.

Using plain Video Vision Transformers (Video ViTs) and combining that with pre-training enables strong Traffic Anomaly Detection (TAD) performance (full details in AITHENA D5.3). It has been found that:

- **Simplifying encoding pipeline:** Advanced pre-training enables simple encoder-only models to match or even surpass the performance of specialised state-of-the-art TAD methods, while also being significantly more efficient.
- **Self- vs. full supervised learning:** Although weakly- and fully supervised pre-training are advantageous on standard benchmarks, they have been found less effective for TAD. Instead, self-supervised Masked Video Modelling (MVM) provides the strongest signal.
- **Domain Adaptation:** Domain-Adaptive Pre-Training (DAPT) on unlabelled driving videos further improves downstream performance, without requiring anomalous examples.

Recommendations

On the use of Vision Language Models (VLMs) combined with a formal ontology:

- Combining the structured way of defining the scene that offers an ontology and the vast knowledge of the VLMs can be beneficial for identifying relevant safety situations. The VLMs can read text in a structured format and understand it, also, they can produce the answer in a specified format. For example, the JSON output is suitable for indicating the system of the presence of something "unusual," which can trigger a fallback or safety protocol.
- Developing a more detailed "reasoning log" captures not just the final output but the step-by-step logic that led to a specific risk alert (e.g., "detected pedestrian at location [x,y] -> pedestrian is in the road area -> risk of collision is high"). This will provide even richer and more robust information for safety-critical events.

On using plain Video Vision Transformers (Video ViTs) and combining that with pre-training enables strong Traffic Anomaly Detection (TAD) performance:

- Use descriptive ontology using VLMs (research above) to improve collision risk dataset even further. It provides a common terminology, which helps in improving quality of annotations of such datasets.

7.2.2. UC2.2 Collision Prediction with Hybrid AI data fusion models

Key insights and lessons learned

- Middleware solutions like the Data Distribution Service (DDS) in Robot Operating System 2 (ROS2) helps connect different software parts, such as object detection, sensor fusion, and decision making, for advanced driver assistance and automated driving. However, just making these parts work together isn't enough in safety-critical areas; they also need to run on time. As these systems get more complex, traditional testing can't cover every possible problem. That's why it's better to guarantee timing during the design phase. Tools can analyse how long each part takes and schedule them, so everything runs safely and on time. Even though GNU/Linux isn't safety certified, it can be used for fast prototyping. By adding a special scheduler to Linux, we can ensure software parts run exactly when they're supposed to, making the system predictable and safe, even when using ROS2 and components from different developers.
- There is a need for high-quality European data sets that adequately cover European ODD-specific elements:

Traffic Light Colour Detection: Robustness across diverse real-world conditions could not be achieved with currently available public datasets alone. Synthetic augmentation, meta-learning strategies, and anchor-free detection methods were critical to improving adaptability and handling rare scenarios such as glare, rain, or high-speed approaches.

Object Orientation & Tracking – Real-world deployment exposed challenges such as motion blur, occlusion, and irregular traffic behaviours, combined with motion-compensated pipelines, enhanced robustness, but also highlighted the need for stronger diagnostic logging to ensure Custom tracking layers combined with motion-compensated pipelines enhanced robustness but also highlighted the need for stronger diagnostic logging for accountability.

Data-Centric Development – Balanced and diverse training data proved more decisive for fairness and robustness than incremental architectural changes. Underrepresented classes (e.g., cyclists, rare traffic signals) consistently drove detection errors, reinforcing the importance of continuous dataset enrichment.

Ethics & Privacy – Despite focusing only on situational awareness and object-level signals, onboard cameras use inevitably raised privacy concerns. Embedding anonymization, strong data-handling policies, and compliance-by-design from the outset proved essential.

- **Safety of systems:** Safety related AI-based system validation in the automotive industry shall follow the newly released ISO 8800 standard: "Safety of systems using artificial intelligence (AI) technologies in road vehicles, ensuring safe integration and operation"

- **Hybrid AI approaches:** Combining model-based and data-driven methods (e.g., physics-based filters with CNN/RNN predictors) enhances robustness, explainability, and domain adaptation in safety-critical contexts such as autonomous driving.
- **Technical maturity gaps:** Data-driven methods often outperform physics-based ones on accuracy metrics, but physics-based approaches remain essential for interpretability, safety, and fallback strategies.
- **Explainability:** Designing models with explainability integrated into their architecture (rather than retrofitted post-hoc) supports trustworthiness, accountability, and regulatory compliance.
- **Scenario variability:** Urban environments introduce complexity due to diverse road users, infrastructures, and unpredictable dynamics. Prediction and reconstruction methods must address uncertainty, noise, and missing data.
- **Data management challenges:** Frequent updates, time-based purging, and memory optimization are critical for maintaining efficient and real-time knowledge in connected mobility systems.

Recommendations

- Strengthen cross-stakeholder collaboration: Facilitate cooperation between research institutions, industry, and regulators to align technological progress with societal needs and legal frameworks.
- Mandate explainability and transparency: Require that AI models in safety-critical domains such as autonomous driving provide interpretable outputs and integrate prior knowledge where applicable.
- Support hybrid R&D: Encourage funding and collaboration for hybrid AI approaches that combine domain knowledge with AI-based solutions to improve safety and generalization.
- Ensure lifecycle oversight: Implement guidelines covering the entire AI lifecycle emphasizing adaptability and continuous validation.
- Establish data governance frameworks: Standardize policies on data retention, purging, and quality assurance to balance efficiency with long-term accountability.
- Promote scenario robustness testing: Develop standardized benchmarks and stress tests to assess model performance under diverse urban scenarios.
- Since AI system performance, robustness, and explainability depend on training, testing, and validation data, these datasets should be stored, versioned, and accessible to authorities and stakeholders. It is recommended to establish a policy for commercially used AI systems.

7.3. UC3: Trustworthy decision making

7.3.1. Key Insights and lessons learned

- **Explanation of driving behaviour:** Decision-making and motion planning represent one of the last intelligent components in the processing pipeline of automated driving systems. The resulting driving behaviour is perceived by occupants and individuals outside the vehicle even

without explicit visualization or similar measures. Insofar as the vehicle behaves in a comprehensible and human-like manner, an explicit explanation may not be necessary. In situations in which driving behaviour is not intuitive, an additional explanation of the behaviour may be helpful in order to increase the trustworthiness of the systems. One example of this could be decisions based on V2X information, as the transmitted information (e.g., the time of a traffic light change or a hidden road user) may not be directly perceivable by passengers.

- **Data for learning-based behaviour- and motion-planning:** The provision of training data for data-driven approaches applied to behaviour and motion planning shows specific challenges. Defining the ground truth poses particular challenges here: the label is usually represented by the desired behaviour, often depicted as a desired trajectory. The subjective desired behaviour is can vary between different groups of people. Furthermore, data from the simulation can only be used for training to a very limited extent. The reason for this is that the neural networks are used to imitate the target behaviour. If simulation data is used for training, the behaviour of the behavioural models from the simulation is learned. In order to learn human-like behaviour, the behavioural models must therefore reflect this accordingly. It might be more effective to acquire data from instructed drivers or use natural trajectory data recorded by external measurement methods (e.g., drones or digital infrastructure).
- **Aspects of vehicle stability in supervised-learning-based approaches:** A particular challenge in the integration of approaches for behaviour- and motion planning based on supervised-learning-based approaches, is the resulting accuracy and robustness with a closed control loop. Approaches based on supervised learning have the problem that they are trained using an open control loop, might result in poor performance when the control loop is closed. The issue was tackled by supplying the network with the current vehicle status at the time of planning as input and augmentation.
- **Software-based driving-guardrails:** AI-based approaches in the context of automated driving should be limited by appropriate boundary conditions at the final processing step in order to obtain a trustworthy overall system. Our approach therefore incorporates a downstream MPC-based algorithm, which examines the AI-based reference trajectories e.g. for collisions with other road users or violations of the drivable space.

7.3.2. Recommendations

- **Behaviour-Dataset development:** Support the development of trajectory datasets representing human-like driving behaviour for training of AI based approaches for behaviour and motion planning. This also requires input data not only labels (which might stem from trajectory datasets) to train the models.
- **Driving Guardrails:** Definition of uniform and necessary driving-guardrails, especially for the integration of AI-based methods for vehicle guidance and motion planning.

7.4. UC4: AI in Traffic Management

In AITHENA, a framework has been developed to assist road authorities to objectively assess AI capabilities in different traffic conditions, situations, and scenarios and assure effectiveness of measures taken and information disseminated to manage traffic on their road network.

7.4.1. Key insights and lessons learned

- **Results from desk study:** Considering different AV penetration rates in a mixed traffic scenario (based on European projects like L3Pilot (Aittoniemi et al. 2023), INFRAMIX (Berrazouane et al. 2019), and Symul8 (Mischinger-Rodziewicz et al. 2024)) different insights were gathered:
 - (a) At low market penetration rates (i.e., 0–40%), AVs often reduce overall efficiency rather than improve it. Safety is fragile, where cautious AV logics tend to increase conflicts during peak conditions. At these early shares, emission effects are limited and not consistently measurable. Low penetration is a “transition” stage where benefits are very sensitive to vehicle settings. Conservative headways depress capacity, but cooperative strategies can already deliver localised gains.
 - (b) Once penetration reaches around 40%, consistent efficiency improvements become more visible. In urban networks, simulations report that at about 50% penetration, average delays reduce by 31%. Safety outcomes improve only if AVs are designed to behave more like attentive human drivers. When they are modelled as “all-knowing” or assertive, conflicts decline; but when modelled as overly cautious, the number of risky interactions remains high. Emission impacts in the medium penetration range (40-70%) vary, where some freeway simulations (at 50% penetration) show increase in CO₂ emissions while others found stabilisation effects that reduced emissions. Medium penetration is the point where systematic gains are possible, but results depend heavily on traffic management strategies and on whether AVs are tuned for cooperation rather than conservatism.
 - (c) At high penetration (70-100%), homogeneous fleets eliminate the behavioural contrast that causes inefficiencies in mixed traffic. In some simulations, congested hotspots show improved travel time, reduced delays and emissions. Safety at full penetration is generally similar to or slightly better than all-human fleets, provided AVs are calibrated correctly. Emissions show clear reductions at full penetration. High penetration brings the most significant local benefits, especially in congested conditions and on critical segments. However, at the scale of entire networks where much of the traffic is not congested, the overall impact may appear modest.
- **AITHENA simulation results:** The simulation results show that the effectiveness of automated vehicles (AVs) is highly dependent on their driving style, the percentage of AVs on the road, and the traffic conditions. Under normal circumstances, cautious AVs create the most significant delays and emissions, especially in congested traffic, while aggressive AVs are most efficient. When an incident occurs in light traffic, trust in traffic management (TM) information can help mitigate delays, but cautious behaviour can surprisingly decrease safety (although it is known that overly cautious driving leads to unsafe conditions). In heavily

congested scenarios with an incident, all AVs worsen traffic, with cautious types performing the worst, and the benefits of trust are nullified by the oversaturated network, leading to a consistent decline in safety and efficiency (full details are provided in D5.3).

It should be noted that these results are specific to the network, AV types, simulation calibration to match real life traffic patterns, traffic demand (Level of Service / level of congestion), and incident logic and cannot be generalised to other operational conditions. The primary aim of this study was to determine the feasibility of the methodology by implementing it for a set (of 156 combinations) of operational conditions.

The developed framework and methodology demonstrate that the approach can be applied across a wide range of operational conditions, with parameters such as AV type, traffic demand (Level of Service / level of congestion), and edge cases tailored to the network under study. This enables a transparent assessment of the impact of AVs on traffic networks using metrics and KPIs grouped into the categories of efficiency, environmental sustainability, and safety. In particular, the analysis highlighted the role of trust in TM information as a critical factor influencing system performance.

- **Parameterisation of AV behaviour is key:** The parameterisation of AV behaviours—such as acceleration, following distance, and lane-changing—was used to define three AV types: cautious, normal, and aggressive. These distinctions proved highly significant, as the results were very sensitive to the chosen parameterisations. Since AVs from different suppliers can vary considerably in their driving behaviours, accurate parameterisation is crucial to ensure realistic representation, especially given the sensitivity of outcomes to these parameters.
- **Mirroring of higher-level AV behaviour:** Just as the parameterisation of lower-level behaviours is essential, higher-level behaviour must also be mirrored in the simulation. How AVs handle incidents in the simulation—whether with or without TM information—depends on their specific characteristics. To what extent do they bend traffic rules, follow surrounding traffic, or possess the necessary algorithms to manage complex situations? The answers to these questions shape the higher-level behaviour of a given AV. Similar to the key insight from UC3 “*Data for learning-based behaviour and motion planning*”, real-world data is required to accurately capture the actual behaviour of AVs in such situations.
- **Real-world behavioural driver data for baseline:** Real-world behaviour in difficult situations is required not only for AVs but also for human-driven vehicles, to establish a baseline. This behaviour must be defined and implemented in the simulation software, as such software does not inherently include the logic to address complex situations. It should be noted that the “desired” behaviour is often subjective and may vary between different groups of people. However, the aim of the UC4 methodology is to reflect actual real-world behaviour. In this context, the desired behaviour is what is observed in practice, regardless of what might be preferred from the perspective of road authorities or formal traffic rules.
- **Real-world traffic data is required for calibration:** To effectively study the impact of introducing AVs into a traffic network, the network and simulation must accurately reflect real-world conditions before the AVs are incorporated. This requires traffic data to calibrate patterns under different operational conditions (e.g., rush hour, Sunday mornings, road works, incidents). In addition, the behaviour of other actors in the traffic system must be understood, including traffic light controller programs and strategic traffic management

rerouting scenarios. In practice, meeting all these requirements can be challenging due to limited data sources and barriers such as complexity or lack of cooperation in obtaining the necessary data.

7.3.2. Recommendations

- **Collect data for edge-cases:** In the list of key insights above, it is stated that the UC4 methodology requires data on how both human-driven vehicles and AVs handle difficult situations (i.e., edge cases). To realistically replicate these behaviours in simulation, real-world data is essential. It is therefore recommended that road authorities collect data on how both human-driven vehicles and AVs respond to such situations.
- **Collect data for disengagements:** To realistically simulate AV behaviour in complex traffic environments, it is essential to collect detailed data on disengagements: moments when the automated system hands control back to a human driver or fails to cope with a situation. Disengagements provide critical insight into the limitations of current AV technologies and how they respond under edge cases, such as unexpected incidents, unusual traffic behaviour, or adverse conditions. These data not only highlight where AVs struggle but also reveal how human drivers handle the same situations, enabling the establishment of a meaningful baseline. By systematically collecting and sharing disengagement data, road authorities can ensure that simulations reflect real-world behaviours more accurately.
- **Require AV providers to share driving behaviour parameters:** Accurate simulation of AV impacts relies heavily on the parameterisation of key driving behaviours, such as acceleration, following distance, and lane-changing behaviour. These parameters determine how an AV drives and the results of simulations have shown to be highly sensitive to variations in these parameters. Since AVs from different providers may differ significantly in their behavioural algorithms, requiring providers to share these driving behaviour parameters is crucial for ensuring realistic and comparable assessments. By mandating transparency in this area, road authorities can build more reliable simulations, support consistent safety evaluations, and better anticipate the effects of AV integration across diverse traffic networks.

8. POLICY RECOMMENDATIONS

This chapter draws together the recommendations outlined in earlier sections, presenting a consolidated set of integrated guidance for the deployment of AI-driven Connected and Cooperative Automated Mobility (CCAM) systems. The recommendations span key areas including human-centred ethical evaluation, data management, AI algorithm development, supporting tools, and testing protocols.

8.1. Human centred ethical evaluation

This section outlines policy recommendations centred on human-focused ethical assessment, incorporating principles such as trust and ethics by design, regulatory harmonisation, safety, testing and certification, as well as transparency and communication.

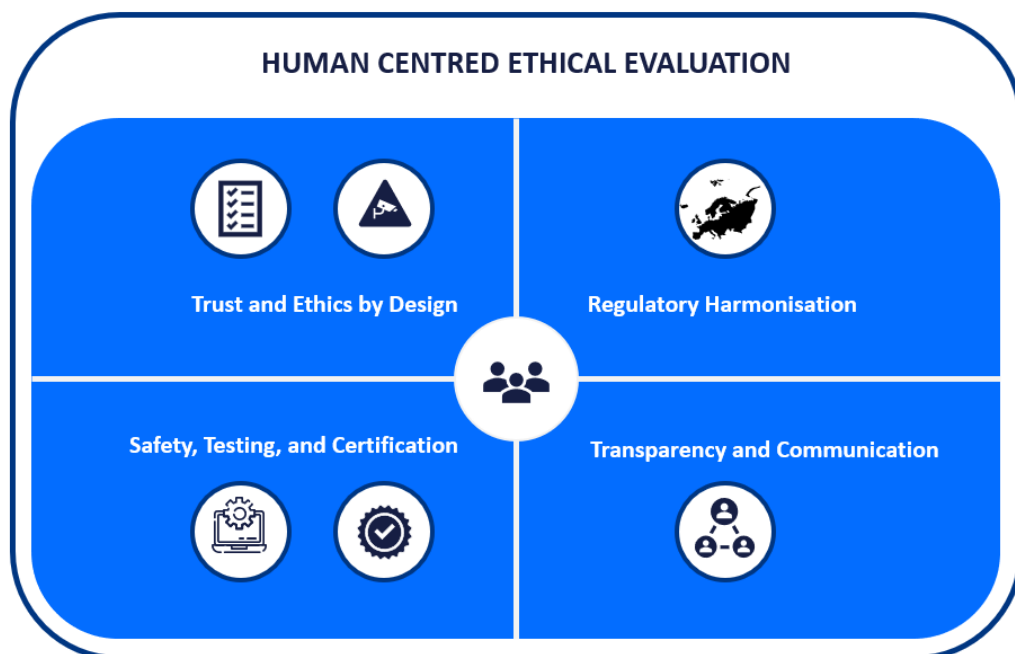


Figure 4: Recommendations for human-centric ethical evaluation

Trust and ethics by design

- **Make 'trust by design' mandatory:** Require a human-centric checklist (i.e., fairness, transparency, accountability, and privacy, as seen in AITHENA's [methodological checklist](#)) to be used throughout development and operations, not just at launch.
- **Protect people's data in practice:** Standardise consent for drivers, passengers, and bystanders; require privacy-by-design measures (minimisation, anonymisation, pseudonymisation, federated options) and appoint DPOs for CCAM operators.

Regulatory harmonisation

- **Coordinate cross-border rules:** Promote mutual recognition of testing and transparency artefacts to reduce regulatory fragmentation across jurisdictions.

Transparency and communication

- **Standardise transparency packs:** Mandate simple, audience-specific disclosures (for users, auditors, and regulators) covering data provenance, model purpose/limits, feature importance, and explanation methods.

Safety, testing, and certification

- **Build testing and simulation into approval:** Require scenario-based tests (incl. uncertainty and edge cases) as part of ethical assurance, not an afterthought.
- **Harden liability & certification links:** Align legal liability with safety certification and type-approval to ensure certified functions are legally usable across Member States.

8.2. Data and life-cycle management

This section offers policy recommendations concentrated on data and life-cycle management. It encompasses measures to strengthen privacy protections and responsible machine learning practices, establish standardised protocols for AI documentation and governance, and encourage strategic dataset stewardship alongside cross-stakeholder collaboration.

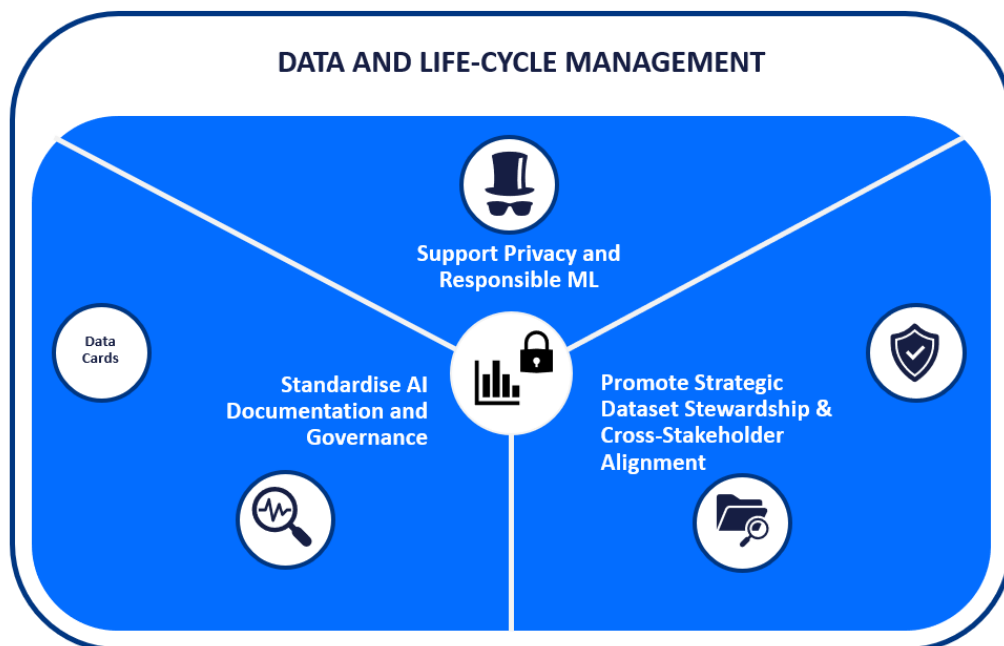


Figure 5: Recommendations for data and life-cycle management

Support privacy and responsible ML

- **Back privacy-preserving ML:** Mandate appropriate anonymisation/pseudonymisation; incentivise secure federated learning where it does not degrade performance; require user-facing privacy controls.

Standardise AI documentation and governance

- **Adopt cards as compliance tools:** Require Data Cards in approvals to document sources, quality, limitations, bias risks, metrics, and caveats.
- **Enforce data governance basics:** Make dataset versioning, lineage, drift detection, and continuous model monitoring compulsory for safety-critical AI.

Promote strategic dataset stewardship and cross-stakeholder alignment

- **Resource high-value datasets:** Identify and maintain priority datasets (incl. European ODD specifics) as shared assets to boost safety and reduce bias.
- **Drive collaboration:** Use common tools (e.g., versioning/visualisation) and shared documentation to align developers, operators, and regulators.

8.3. Explainable Continuous Model Development

This section presents policy recommendations focused on the development of explainable continuous models. It includes provisions to require the design of AI models that are both explainable and transparent, establishes standards for model fusion and rare-event validation, and calls for impact assessments prior to deployment.

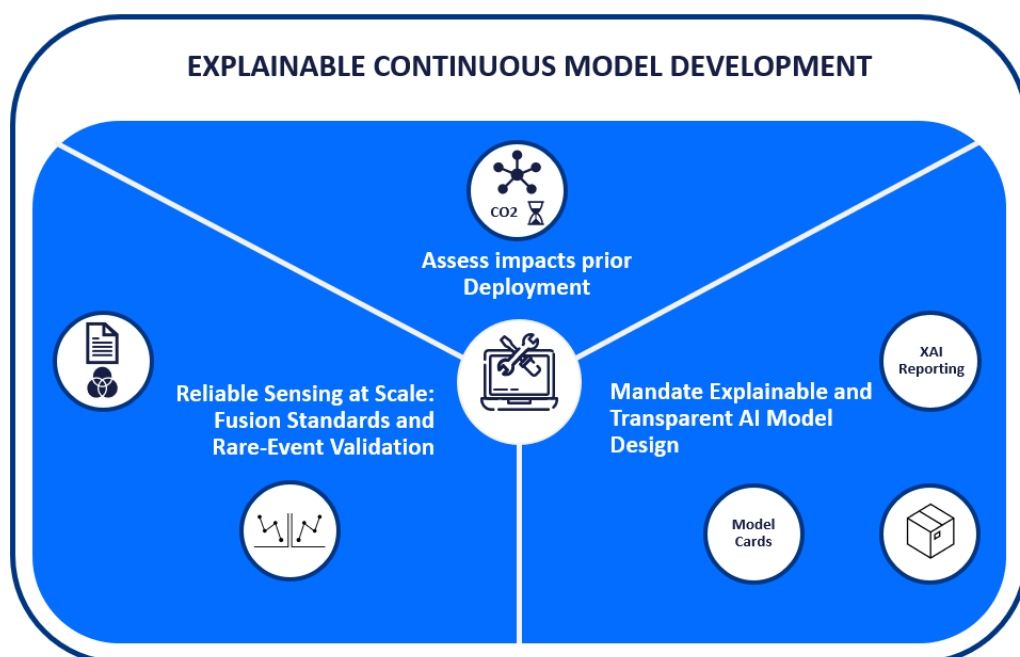


Figure 6: Recommendations for Explainable Continuous Model Development

Mandate Explainable and Transparent AI Model Design

- **Make Model Cards mandatory for every deployed AI model:** Use them to standardise lifecycle transparency from design to updates.
- **Require explainability where safety matters:** Set minimum XAI reporting (local + global explanations, uncertainty outputs) for perception/prediction/planning.

- **Promote hybrid & white-box approaches:** Encourage designs that blend model-based and data-driven methods and integrate explainability into the architecture (not just post-hoc).

Reliable Sensing at Scale: Fusion Standards and Rare-Event Validation

- **Standardise multi-sensor fusion expectations:** Publish guidance for robust LiDAR–radar–camera fusion and require robustness under adverse conditions.
- **Invest in edge-case readiness:** Fund real + synthetic data to cover rare events and provide evidence of performance under those conditions before scale-up.

Assess impacts prior deployment

- **Simulate network-level effects before deployment:** Require city/network-scale assessments (e.g., mixed traffic, varied AV behaviours, incident response) to understand cumulative effects for safety, congestion, and emissions.

8.4. Tools, Testing, Deployment & Validation

This section presents policy recommendations focused on the tools, testing, deployment & validation. It includes provisions to require robustness and runtime transparency, safety architecture and integration, data foundations and sharing and tooling, facilities and policy alignment.

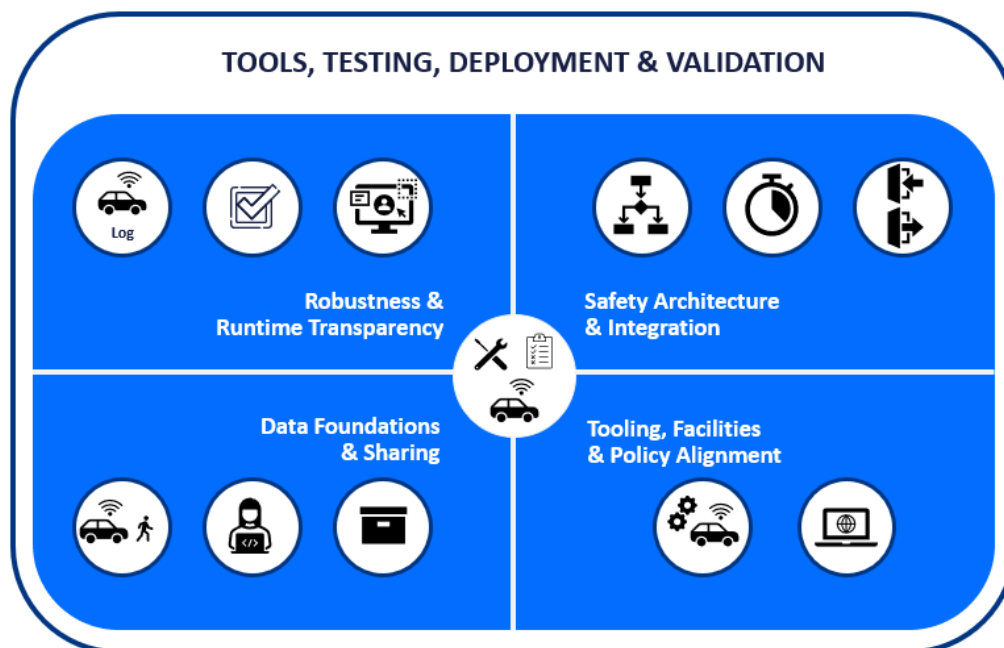


Figure 7: Recommendations for Tools, Testing, Deployment, and Validation

Robustness & Runtime Transparency

- **Establish standard tests beyond benchmarks:** This includes (but not limited to) taking into consideration sensor failures/misalignments, bad weather, OOD objects/situations, and uncertainty/competence reporting at runtime.

- **Log what the AI ‘thought’:** Require AVs to log uncertainty and XAI signals (e.g., saliency/attention, confidence) for incident investigation and continuous validation.
- **Human-system interaction under failure:** Standardise how automated systems communicate system limits and handover needs to users.

Safety Architecture & Integration

- **Guardrails on decision-making:** Define uniform software ‘driving guardrails’ (e.g., downstream safety checks like MPC collision screening) to bound AI planners.
- **Time-determinism in integration:** When using ROS2/DDS and mixed vendors, require timing analysis/scheduling so safety functions run predictably.
- **Lifecycle oversight:** Tie virtual + track + real-world testing to clear entry/exit criteria; mandate in-service monitoring for drift and trigger rules for model updates.

Data Foundations & Sharing

- **Collect edge-case and disengagement data:** Calibrate realistic behaviour of humans and AVs.
- **Share key behavioural parameters:** Require AV providers to share key behavioural parameters (headways, lane-change logic, acceleration) for fair, comparable simulations.
- **Store and version datasets:** Ensure accessibility of training/testing/validation data to authorities.

Tooling, Facilities & Policy Alignment

- **Physical layer:** Use modular vehicle platforms for the ease of swapping sensors and algorithms without rebuilding the whole vehicle. Strengthen redundancy in localisation combining GNSS/INS with vision-based methods, so that positioning remains reliable even when GPS is weak or unavailable. Standardise sensor calibration protocols across partners to ensure that measurements are comparable and integration is smoother. Encourage automated data pipelines (e.g., unified logging formats and automatic upload/processing) so test data flows reliably from the vehicle to shared repositories, reducing manual effort and errors and enabling quicker analysis and re-use.
- **Virtual layer:** Build high-fidelity digital test tracks and vehicle models so that what you learn in simulation matches real-world behaviour more closely. Distribute simulation capabilities that can handle complex communication delays and many interacting agents. Encourage hybrid sensor emulation, mixing real and simulated sensor feeds to stress-test perception and planning before field trials. This should be supported by reusable scenario libraries.

9. CONCLUSION

This report has identified several foundational requirements and best practices for advancing Connected, Cooperative, and Automated Mobility (CCAM) with a focus on artificial intelligence integration, testing, and regulatory oversight. Looking ahead, the evolution of AI in CCAM hinges on a strong alignment between technological innovation and robust, adaptive regulatory frameworks. As AI-driven systems become more prevalent, their capacity to learn and adapt continuously will demand continuous oversight, transparent data sharing, and harmonised standards across industry stakeholders. Policymakers must keep pace with these technological advances by mandating lifecycle management, ongoing performance monitoring, and interoperable safety benchmarks.

Ultimately, building public trust in automated mobility will depend on collaborative efforts to ensure fairness, accountability, transparency and privacy, making CCAM not only smarter but safer and more inclusive for all road users.

REFERENCES

- Aittoniemi, E., Itkonen, T., Innamaa, S. (2023). Travel time, delay and CO₂ impacts of SAE L₃ driving automation of passenger cars on the European motorway network. *European Journal of Transport and Infrastructure Research*, 23(1), 1-29.
- European Commission, "Commission Implementing Regulation (EU) 2022/1426 on uniform procedures and technical specifications for type-approval of automated driving systems (ADS)," *Official Journal of the European Union*, L 221, 26 Aug 2022. [Online]. Available: https://eur-lex.europa.eu/eli/reg_impl/2022/1426/oj/eng
- European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 on harmonised rules for artificial intelligence (Artificial Intelligence Act)" *Official Journal of the European Union*, published 12 July 2024. [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
- European Parliament and Council of the European Union, "Regulation (EU) 2018/858 on the approval and market surveillance of motor vehicles and their trailers" *Official Journal of the European Union*, L 151, 14 Jun 2018, pp. 1–218. [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2018/858/oj/eng>
- European Parliament and Council of the European Union, "Regulation (EU) 2019/2144 on type-approval requirements for motor vehicles and their safety and environmental performance," *Official Journal of the European Union*, L 325, 16 Dec. 2019. [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2019/2144/oj/eng>
- International Organization for Standardization, "ISO 21448:2022: Road vehicles – Safety of the intended functionality", 2022.
- International Organization for Standardization, "ISO 26262-6:2018: Road vehicles – Functional safety – Part 6: Product development at the software level", 2018.
- International Organization for Standardization, "ISO/PAS 8800:2024: Road vehicles – Safety and artificial intelligence", 2024.
- International Organization for Standardization, "ISO/TS 5083:2025: Road vehicles – Safety for automated driving systems – Design, verification and validation", 2025.
- M. Berrazouane, K. Tong, S. Solmaz, M. Kiers and J. Erhart, "Analysis and Initial Observations on Varying Penetration Rates of Automated Vehicles in Mixed Traffic Flow utilizing SUMO," 2019 IEEE International Conference on Connected Vehicles and Expo (ICCVE), Graz, Austria, 2019, pp. 1-7, doi: 10.1109/ICCVE45908.2019.8965065
- Mischinger-Rodziewicz, Marlies and Hofinger, Felix and Haberl, Michael and Fellendorf, Martin (2024), Non-Compliant Behaviour of Automated Vehicles in a Mixed Traffic Environment. Available at SSRN: <https://ssrn.com/abstract=4893379> or <http://dx.doi.org/10.2139/ssrn.4893379>
- United Nations Economic Commission for Europe, "ECE-TRANS-WP.29-2023-44-r.1e: New Assessment/Test Method for Automated Driving – Guidelines for Validating Automated Driving

System (ADS),” Inland Transport Committee, World Forum for the Harmonization of Vehicle Regulations (WP.29), 7 June 2023. [Online]. Available: [GRVA-16-39](#)

United Nations Economic Commission for Europe, “GRVA-13-04r1e: Considerations on the Use of Artificial Intelligence in the Context of the New Assessment/Test Method (NATM),” Working Party on Automated/Autonomous and Connected Vehicles (GRVA), 22 May 2023. [Online]. Available: [UNECE and automated vehicles](#)

United Nations Economic Commission for Europe, “UN Regulation No. 155 on the cybersecurity and cyber security management system for vehicles,” *Official Journal of the European Union*, L 82, 9 Mar. 2021. [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2021/387/oj/eng>

United Nations Economic Commission for Europe, “UN Regulation No. 156 on software update and software update management system for vehicles,” *Official Journal of the European Union*, L 82, 9 Mar. 2021. [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2021/388/oj/eng>

United Nations Economic Commission for Europe, “UN Regulation No. 157 on automated lane keeping systems (ALKS),” *Official Journal of the European Union*, L 82, 9 Mar. 2021. [Online]. Available: <https://unece.org/transport/documents/2021/03/standards/un-regulation-no-157-automated-lane-keeping-systems-alks>

Verband der Automobilindustrie e.V. (VDA), Automotive SPICE® Process Reference and Assessment Model, Version 4.0, 2023, 2023.