Towards a meet-in-the middle approach for Trustworthy AI for CCAM

Karla Quintero, Clément Arlotti, Atia Cortés, Stathis Antoniou

Towards a meet-in-the middle approach for Trustworthy AI for CCAM

Karla QUINTERO[a,1], Clément ARLOTTI[a], Atia CORTES[b], Stathis ANTONIOU[c]
[a] *Institute of Research and Technology SystemX, France*
[b] *Barcelona Supercomputing Center, Spain*
[a] INLECOM Innovation, Greece

Abstract. This work tackles the problem of building trustworthy AI for the automotive industry in a context in which generic guidelines have already been proposed yet their instantiation is far from straightforward. The following work presents a first iteration of a methodology for developing trustworthy AI in CCAM (Connected, Cooperative Autonomous Mobility) applications as a meet-in-the-middle approach integrating generic European ethics guidelines (top-down) as well as leveraging the scenario approach (bottom-up) as a well-known practice in the automotive field. The result is a first version of application of the trustworthiness criteria into a use case of AI-enhanced ADAS and a related scenario subset. The premise is that in order to truly develop trustworthy AI, trustworthiness criteria are necessary but must be coupled with solid practices in the field and systems of reference in order to ensure integration of ongoing and proven engineering processes to the new challenges and opportunities linked to the development cycle of AI-based systems.

Keywords. Trustworthy AI, automated vehicles, scenario-based testing, ML models.

1. Introduction

Artificial Intelligence (AI), and in particular machine learning (ML), are expected to play a key role in the deployment of Connected, Cooperative, Autonomous Mobility (CCAM). However, this opportunity comes with several ethical, legal, social and technical challenges (e.g. transparency, fairness, privacy-preserving, safety, sustainability, accountability, among others) that should be addressed to enhance trustworthiness across the ML-based system's life cycle. Namely, most of the current challenges regarding trustworthiness are focused on ML-based AI (connectionist AI). Conversely, rule-based (symbolic) AI already offers various guarantees and trustworthiness-related properties (e.g. deterministic outputs, explanations in the form of rules, abductive reasoning, among others) but does not leverage the predictive power and flexibility of the data driven approach; therefore, the underlying challenge is to bring ML-based systems to this level of trustworthiness, so as to exploit their full potential in CCAM and other critical systems.

The core of this work is the methodology from the AI4CCAM project that integrates the scenario approach as a classical method and tool in the transport sector, AI

[1] Corresponding Author, Institute of Research and Technology SystemX, 2 Boulevard Thomas Gobert, 91120 Palaiseau, France; E-mail: karla.quintero@irt-systemx.fr.

trustworthiness criteria as identified at a European level (e.g. [1], [2]), and results from the multi-sector Confiance.ai program. The latter has tackled over the past 3 years: trustworthiness-related attributes, methods, scores, models and components for safety-critical systems based on synergies with pre-existing development methodologies.

The contribution herein hinges around the scenario approach, to prescribe and assess Automated Vehicle (AV) capabilities and its application together with AI-related ethical guidelines and criteria which are proposed, to this date, in a broader scope. We develop a method and give an example of the instantiation of these criteria in coherence with a specific use case that addresses an AI-enhanced Advanced Driving Assistance System (ADAS). To our knowledge the instantiation of these criteria is not straightforward. Moreover, it can be open to misinterpretation.

The purpose of our work is thus to provide an unambiguous approach, adapted to the development of ML-based models, in coherence with the existing reference system (i.e. in the CCAM scope) to assess trustworthiness-related criteria in all phases on the development cycle so that:

- ML-based models are trustworthy by design,
- ML-based systems are properly traced and documented for assessment or audit as criteria are evaluated in the pertinent phases,
- ML-based systems are properly traced for iterations, corrections, and upgrades for new versions regarding trustworthiness criteria,
- Scenario-based testing, widely used in the automotive field, is still used in coherence with ongoing industrial processes as well as in alignment with the integration of ML-based models from the design phase onward.

The paper is organized as follows: section 2 covers the context and related work, section 3 presents a methodology for trustworthy AI in CCAM from a high-level perspective, section 4 shows the instantiation of one phase of the methodology with respect to a set of trustworthiness criteria for one specific use case, and section 5 covers the perspectives and future work.

2. Context and Previous Work

The scenario approach or scenario-based testing stands as a state-of-the-art methodological enabler to provide means for verification and testing of automated vehicles, if not fully at least to the extent to which conventional testing is no longer feasible due to the number of situations to cover (see [3] and [4]). Moreover, the integration of AI-based functions only complexifies the endeavor since potential hazards can arise as sensors or AI-models can be exposed to disturbances that are not to be addressed by the driver. Extensive work in the automotive field currently focuses on the best strategies to address the combinatorial explosion of possible scenarios. In this sense, ‘proper’ coverage of the ODD is paramount; the reader can refer to [5] for an ODD-driven coverage for safety argumentation of AVs.

The re-allocation of responsibility from the driver to the system in situations that were not modeled, verified, and tested systematically before is a fundamental challenge for trustworthy AI. As stated in [6] ”If the concept of ‘Trustworthy AI’ is kept being used, we risk attributing responsibilities to agents who cannot be held responsible, and consequently, deteriorate social structures which regard accountability and liability”. Striving to avoid this pitfall thus raises the need for a transparent, ethics-informed,

analytical methodology, allowing each step and actor of the design process to be identified as accountable in the overall reliability chain of the system operation.

On the technological front of the subject of Trustworthy AI, a main pillar of the work herein is constituted by the efforts of the French program Confiance.ai [7]. This multi-sector and multi-disciplinary program, with over 50 industrial partners from industry and academia, is tackling the subject of engineering trustworthy AI and the integration of trustworthy ML-components to pre-existing and widely deployed industrial processes. The work in the Confiance.ai program covers, among others, trustworthiness criteria that find a vast common ground with those defined at a European level defined and used further in this paper.

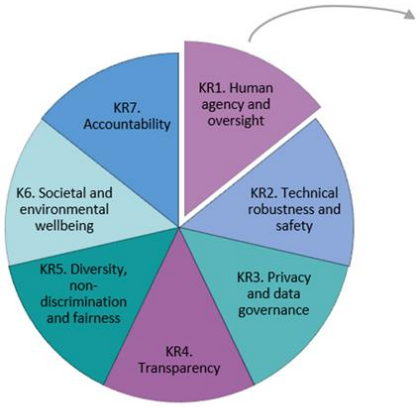
European guidelines are at vanguard pushing toward responsible design, implementation and deployment as will be described in this section. The definition of a European ethical and regulatory framework for Artificial Intelligence has been essential to put forward the concept of trustworthiness and human-centric AI and to highlight the need to involve experts from other disciplines coming from social sciences and humanities among others. Over the last years, hundreds of guidelines, codes of conduct or standards that define an ethical framework for AI have been released worldwide. Some documents have been fundamental for the purpose of this work and of the AI4CCAM project, in particular 1) the Ethics Guidelines for Trustworthy AI [1] that was produced by the High-Level Expert Group on Artificial Intelligence (HLEG-AI) set up by the European Commission; and 2) the Trustworthy Autonomous Vehicles report [4] from the Joint Research Center of the European Commission.

The ethical guidelines present the following framework for Trustworthy AI, based on three layers:

- The first layer presents four ethical principles that define the foundation of the Trustworthy AI framework: respect for human autonomy, prevention of harm, fairness, and explicability. The former three principles are clearly related to bioethical principles and fundamental rights, while the latter regards the process to explain any decision related to the development, deployment and use of the AI-based system.
- The second layer introduces seven ethical key requirements to implement these ethical principles: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental wellbeing; and accountability. These requirements need to be evaluated throughout the AI system's life cycle.
- The third layer involves methods to operationalize the key requirements. Multiple techniques and methodologies have been proposed over the last years to assess the trustworthiness of AI/ML-based systems – partially or as a whole-, however there is yet no standard procedure to do it. Along with the guidelines, HLEG-AI released the Assessment List on Trustworthy AI, which aims to be used as a self-evaluation process and contains a set of questions to assess each of the seven requirements [8].

In the context of Connected and Automated Vehicles (CAV), the work in [2] is key since it transposes the Trustworthy AI Framework in the scope of these particular systems and throughout their entire life cycle. This report identifies seventy assessment criteria for the domain of AVs associated to each of the seven EU key requirements. Moreover, it classifies criteria as critical/short-term, important/mid-term and impact/long-term in relation to the level of relevance and urgency of evaluation. As an

example, one key requirement developed in this paper is human agency and oversight, and the prioritized criteria in the report are as presented in Figure 1. These are then key attention points that should be analyzed, characterized and documented when ensuring trustworthiness of AI-based systems in automated vehicles. For details on the criteria identified for all 7 key requirements, namely for connected and automated vehicles, the reader can refer to [2]. Figure 1 illustrates the criteria prioritized in this report for the first key requirement: "Human agency and oversight"; for other requirements; specific criteria are identified, as an example, for key requirement 2 "Technical robustness and safety" 22 criteria are prioritized in the domains of: resilience to attack and security, general safety, accuracy and reliability and fallback plans and reproducibility.



CRITERIA

Human Agency and Autonomy	Human Oversight
CR1.1: Affects humans or society	CR1.7: Self-learning or autonomous / Human-in-the-loop / Human-in-Command
CR1.2: Confusion as to wether the interaction is with a human or an AI	CR1.8: Training on how to exercise oversight
CR1.3: Overreliance	CR1.9: Detection and response mechanisms for undesirable adverse effects
CR1.4: Unintended and undesirable interference with end-user decision-making	CR1.10: Stop button
CR1.5: Simulation of social interaction	CR1.11: Oversight and control of the self-learning or autonomous nature of the AI system
CR1.6: Risk of attachment, addiction and user behaviour manipulation	

Figure 1. Trustworthiness criteria for key requirement1: human agency and oversight [2]

To this date and to our knowledge there is no unequivocal process to specify the methodology through which these criteria should be studied and implemented in the CCAM field and this work proposes one such method in the scope of simulation.
Current European initiatives addressing this challenge include the AI4CCAM project [9], which has produced the results in this work that will continue until end of 2025. This project addresses trustworthiness of AI in the context of CCAM and encompasses the subject of trustworthiness of AI-models for AI-models performing VRU trajectory prediction among others.

Finally, within the scope of AI-trustworthiness in safety-critical systems, beyond the ethical scope, very recent work published in [10] addresses a compendium on trustworthiness attributes, the underlying issue of the integration of conflicting attributes, the role of multi-criteria decision making (MCDA) and an approach based on a metamodel of attributes allowing to tackle conflicts and commensurability in an understandable manner for stakeholders. For the purposes of the work herein, it is considered, as stated in [10] that: "In addition to measures and processes, various techniques and methodologies such as testing, evaluation, and validation of the system's performance against specified criteria, expert review, and stakeholder participation are required for trustworthiness assessment in AI-based critical systems". That work, is considered key for further developments beyond the work herein.

As aforementioned, numerous initiatives have marked the evolution of how trustworthiness should be assessed, some of them through different lenses (e.g. technological, ethical) and most of them implying self-assessment. Ethical guidelines have led to specific criteria definitions which put the responsibility of the priority or weight of each criteria (where conflicting ones are identified) in the control of the system provider. This has motivated the development of a regulation, the AI Act [11], aiming at better allocating responsibility and attention points on trustworthiness depending on the potential risks induces by the system.

3. Methodology for Trustworthy AI for CCAM

The contribution of this paper is related to the methodology allowing to integrate:

1. European initiatives and subsequent criteria related to AI trustworthiness, specially related to ethical guidelines,
2. the well-known scenario approach to model situations to be encountered by the System Under Test (SUT) including the aforementioned trustworthiness criteria when possible at that stage, and
3. the overall pipeline for trustworthy AI for safety-critical systems based on intermediate results of the Confiance.ai program.

This meet-in-the middle approach is presented in Figure 2.

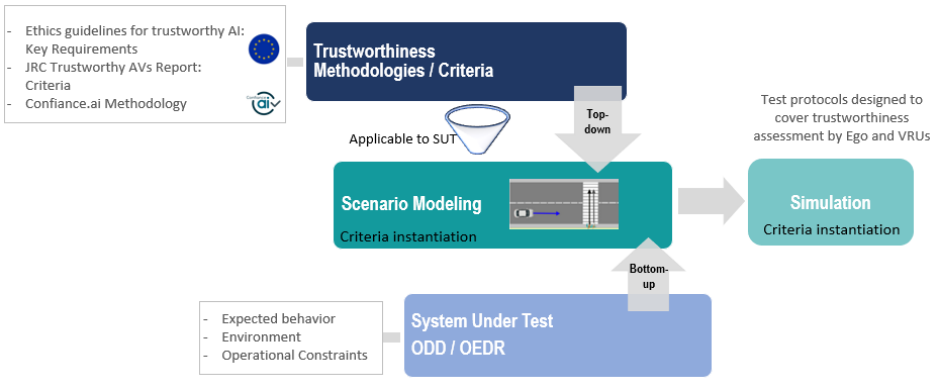


Figure 2. Meet-in-the-middle approach for trustworthy AI assessment on the CCAM scope

The scenario approach is a technical-oriented, bottom-up framework that lacks embedded ethical guidelines. Conversely, European ethical requirements remain elusive on how to implement them in the life-cycle of a real-world system (i.e. from design, to deployment, through development and validation). The meet-in-the-middle approach presented on **Figure** 2 shows the necessary convergence of these 2 currents with an envisaged implementation on simulation.

The proposed methodology herein is based on a macro decomposition of phases in a pipeline to ensure trustworthiness when developing a given AI-based system for CCAM, inspired from the *confiance.ai* program [12] and the intermediate results in 2022. The pipeline, even though designed for multi-sector industrial applications, allows however circumscribing specific activities in the project at a high-level. Trustworthiness

criteria are addressed for each one of the phases in the pipeline on the basis of state-of-the-art European developments as those published by the Joint Research Centre report on trustworthy autonomous vehicles in 2021, see [2]. The approach is to decompose activities related to the development of an AI-enhanced function into 4 high-level crucial phases that allow for a trustworthiness-driven approach from design. This approach accounts for scenario modeling as one key domain-specific practices, and each one of these major phases is to be analyzed through the lens of the 7 key requirements and derived criteria as defined in [2].

The methodology is presented in Figure 3, and a first application has been performed in the AI4CCAM project for each phase of the pipeline with the focus on one use case for AI enhanced ADAS (Advanced Driving Assistance Systems). The scenario approach is used to feed the methodology from the problem specification phase and it is applied for modelling one of the use cases in the project, for which preliminary results are presented.

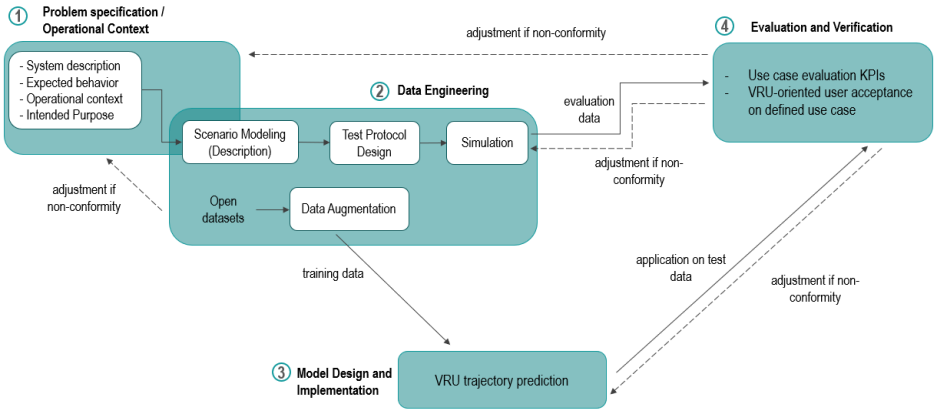


Figure 3. AI4CCAM Pipeline for Trustworthy AI by design

The process presented in Figure 3 can be understood as follows.

Four main phases are identified in a high-level pipeline in order to develop AI models for a given purpose (maintenance being considered as outside the scope of this work). The objective is to address trustworthiness implications from the very beginning of the cycle and specific to each phase. *Phase 1* covers the problem specification, this includes describing the problem to solve, the proposed solution, its operational context and, therefore, in this particular automotive application, the scenarios modeling and the expected behavior of the system. *Phase 2* includes data engineering processes that start with the scenarios that are prescriptive of the test campaigns that will be deployed in simulation. This creates datasets that will be used for evaluating the AI (ML-based) models that will be developed in *Phase 3*. In parallel, in phase 2, datasets are built for ML model training which in the project's scope is based on open datasets and data augmentation. Subsequently, models are designed and trained in phase 3. Finally, *Phase 4* tackles the evaluation of the models based on:

a. key performance indicators for what was specified for the ego vehicle using the AI-enhanced ADAS in phase 1 through the trustworthiness lens (i.e. instantiating applicable trustworthiness criteria and checking points), the data

that was generated on phase 2 also through this lens, and the models that were designed and trained in phase 3 also with these considerations, and

b. VRU (Vulnerable Road User) oriented user acceptance trials that will be set up to present to users, through a virtual reality environment, scenarios equivalent to the ones simulated and used to test the ego's response. In this case the purpose is to collect the reactions (subjective and physiological) from the VRUs in order to analyze correlations with the specified dynamic parameters for the ego vehicle.

The pipeline in Figure 3 aims at addressing, within a simulation testing scope, trustworthiness of AI by design from the problem description phase, going through model development and up until the evaluation phase. It is a macro-decomposition of large phases that allows nonetheless placing project activities as they would be in a conventional engineering cycle and such decomposition enables the application of the criteria proposed in [2] for autonomous vehicles, this without restriction of enrichment in the future. Only high-level phases of interest for the project are shown in Figure 3, and industrial deployment would entail addressing many others in the development cycle, such as conformity with regulation for the reference system, system requirements (from the top-down perspective in the development cycle) and integration, verification, validation, assurance cases (from the bottom-up perspective).

For each phase in the pipeline, all 7 European ethical key requirements and the associated evaluation criteria for trustworthiness are to be addressed. If not applicable, the recommendation is to leave the suitable trace in the documentation and when it does apply, then for the context in this work, each are declined in:

- specific indicators to be computed in the scope of simulation data generation and/or,
- subjective appreciations from users if the criteria are not judged instantiable through objective numeric interpretation (in which case further results in the project will attest on the need to enrich or modify the criteria), and/or
- proper documentation from system providers if deemed to be addressed through simple OEM or system provider internal traceability on design and implementation choices

In the following, the 4 main phases tackled in the project (and depicted in Figure 3) at this point are described.

Phase 1: Problem Specification: which includes the Operational Design Domain (ODD) of the system, its operational context and the general problem that it aims at solving and through which means. This therefore should cover the Intended Purpose as proposed in the AI Act which should have its due impact in the following phases in the pipeline. All of the aforementioned elements entail trustworthiness aspects and attributes. Therefore, in a general manner, through the trustworthiness lens, in this phase of the pipeline the following high-level question is tackled: *Does the system specification itself inherently introduce biases or violations of the requirements of the trustworthy AI framework?* and to answer it, the proper analysis of the proposed criteria should be deployed. This is, addressing all 7 key requirements proposed in the ethical guidelines in [1] and minimally, the criteria identified in [2] as critical for autonomous vehicles. A focus on this phase of the pipeline is presented in section 4.

For the purposes of this work it is to be noted that in this phase, scenario description is tackled (see Figure 3) which is key and a consolidated practice in the ADAS domain.

Here it has been considered that it is also relevant in the data engineering process since it covers designing how simulation datasets will be obtained for training ML-models.

Phase 2 - Data engineering: Which includes every process involving data prior to model design, e.g. collection, preparation, and segregation, compliance with data privacy regulations (e.g. GDPR). Among others, data should be sizable, accessible, understandable, reliable, and usable. These notions are complex and their definition and sufficient justification are debatable depending on the field. Broadly stated, in the scope of trustworthy AI, data engineering processes should be performed in a way that maximum reduction of biases is ensured as well as the related justification and documentation. Herein, for purposes of simplicity, data engineering is circumscribed to a phase in the aim of stressing the need to focus on these aspects. In practice, issues related to data engineering actually impact all phases in the design, development, testing, evaluation and potentially maintenance processes.

Phase 3 - AI Model Design and Development: This phase involves the training of the model as well as its refinement during testing. In this phase, emphasis should be put in the quality and representativeness of training and testing datasets, as well as identification and quantification of their biases. The same analysis should be made for the choice of hyperparameters and how they impact both model performances and its ability to meet the retained ethical criteria.

In the project, the specificity of sub-phases in the development cycle of the AI models is not addressed. A broader view is studied; as an example, the architecture of the system to be developed, whether it is the ADAS or the AI model assisting the ADAS is not developed and studied as a white box in the project. Partners whom are system providers bring and integrate their products and the high-level assessment is applied.

Phase 4 - Evaluation: This phase involves testing the model, and send it back to necessary adjustments in the previous phase when non-compliant with the initial problem specification (i.e. ability to address the specified operational context, behavior coherent with the system's intended purpose, among others). The lack of bias and representativeness of the KPIs chosen to perform the evaluation and verification are then crucial.

A fifth phase proposed in the pipeline is not addressed in the methodology at the moment of writing of this document and it involves the implementation, documentation, deployment and maintenance. It is however worth mentioning its importance in subsequent stages in order to guarantee transparency and enable accountability. This phase is considered out of scope at present time given the context and reach of the project and related use cases.

4. Focus on Phase 1: "Problem Specification"

This section describes concretely the approach deployed for all 4 phases in the pipeline, specifically for the Problem Specification Phase. As aforementioned, per phase, all key requirements are surveyed to pinpoint trustworthiness aspects relevant to the use case and all proposed criteria in [2] are instantiated to evaluate whether the criteria are met or not. The use case is an AI-enhanced ADAS that tackles trajectory prediction for VRUs in urban scenarios.

Table 1. Application of Key Requirement 1 'Human Agency and Oversight' and the derived criteria in [2] to Phase 1 (Problem Specification) in the methodology

Criteria for KR1	**Application on use case: AI-enhanced ADAS**	**Instantiation (Indicator or Support Documentation or other)**
CR1.1 Affects humans or society	Description of the Intended Purpose of the system and hence how it affects humans and society. Prior to development, the system should be specified so that safety parameters are respected, this includes but is not limited to: - take over maneuver (and associated features: proper timing when requested by the AV, seamlessness, proper HMI, seamless disengagement when requested by the user, among others) - safe distances to VRUs -proper HMI for activation/deactivation, oversight on state of operation	- Documentation on Intended Purpose - Documentation on takeover conditions - In simulation, analysis on situations where take over is requested by the vehicle. Given the dynamic conditions identified on these test cases, documentation on distributions and expected behavior - Statistical representation of distances from ego to VRUs in the test campaigns - Presence of HMI module allowing the oversight of the state of the function (active or not) - Presence of mechanism to stop the function if needed
CR1.2 Confusion as to whether the interaction is with a human or an AI	N/A - out of scope in the project. No implementation expected for the VRU to tell the difference between the 2 modes.	
CR1.3: Overreliance	To be foreseen from the driver's perspective as well as from the VRU. To be considered in the problem specification phase to properly state the design requirements to ensure clear and well-defined conditions and limitations of use for the system. These should also translate into proper communication of these conditions though HMI specification.	- Presence of HMI module allowing the oversight of the state of the function (active or not), as defined in CR1.1 - Clear documentation on operating conditions of the function and limitations, understandable to the final user
CR1.4 Unintended and undesirable interference with end-user decision-making	N/A – out of scope since the context is on simulation	
CR1.5: Simulation of social interaction	In the problem specification, the provider shall consider the operational context which includes presence and interaction of VRUs. This implies addressing variability of behaviors of VRUs, including a sense of uncertainty and therefore the underlying measures to address it weather it is through integration of the models or constraining the use of the system to reliable conditions in the sense of interactions of users around the AV.	Addressing test cases as the following to assess the pass/fail criteria - 'Erratic' movement for VRUs surrounding the ego vehicle - VRU Crowds surrounding, next to, and in front of ego. - Rapid crowd movement - Group dissociation in front of ego: one big target becomes several targets at different angles.
CR1.6 Risk of attachment, addiction and user behavior manipulation	N/A to this function	

CR1.7 Self-learning or autonomous / Human-in-the-Loop / Human-on-the-Loop / Human-in-Command	proper documentation to driver of the vehicle	Proper documentation to driver of the vehicle
CR1.8 Training on how to exercise oversight	Simple, proper documentation to end user. The driver should be aware of the existence of the function and its default mode. The question of awareness of the function for VRUs surrounding the vehicle is still a research subject in itself since new, challenging and often unsafe behaviors can emerge from knowing the function is active. No recommendation is given for now regarding oversight from VRUs.	Simple, proper documentation to ego driver.
CR1.9: detection and response mechanisms for undesirable adverse effects	The operational context induces risk assessment on potential undesirable adverse effects and the expected response of the system to these should be specified. Some of these can include: adverse environmental conditions, occlusion, misuse from VRU potentially due to overreliance (CR1.3) or malicious intent.	Besides due OEDR (Object and Event Detection and Response) specification, test campaigns should include scenarios with: occlusion of VRUs, variability of adverse environmental conditions, whether they are in the ODD (therefore managed by the system) or out of it, in which case the system should request the drive to take over.
CR1.10: stop button	Possibility to ergonomically deactivate the function if the driver sees it fit.	Stop button and visual confirmation of deactivation on HMI
CR1.11 Oversight and control of the self-learning or autonomous nature of the AI system	Specific documentation for driver. No recommendations yet for VRUs for the same reasons stated in CR1.8.	Specific documentation for driver

The first phase of the methodology is the problem specification which includes the operational context. For this phase, all 7 key requirements and related criteria should be covered and Table 1 synthesizes the results for the first key requirement: Human Agency and Oversight. All aforementioned suggestions of application of the criteria to the AI-enhanced ADAS use case in the project are considered the minimal necessary and do not exclude further analyses to be performed.

As depicted in the methodology on Figure 3, core elements of the problem specification are represented through descriptive scenarios in the project in order to prescribe simulations through which the trustworthiness criteria (as the ones mentioned in Table 1 for example) can be assessed. An example of a descriptive scenario addressed in the project is the one depicted on Figure 4 in the MOSAR Scenario Manager. This logical scenario (meaning parameter ranges are specified for testing) is the subject of the first simulation campaign in the project and involves 3 actors: 2 pedestrians crossing the street on a crosswalk as the ego approaches. These parameter ranges allow test strategies to be established and simulation campaigns to be designed. Results from simulations per specific test case can be retrieved and integrated back into the descriptive logical scenario in order to perform statistical analysis on the obtained results.

In Figure 4, a logical scenario is described as a storyboard (i.e. scene sequence) in which each scene involves actors and environment parameters. For actors, the default parameters that need to be described per scene are as depicted in Figure 5, and similar parameters are considered for each pedestrian. Additionally, weather default parameters include: luminosity, light density, temperature, rain, fog visibility, wind (velocity, direction), nebulosity, snow, hail, and smoke visibility. These parameters are the base for testing a specific AI model since they should account for the influencing factors depending on the sensors and then subsequent data processing.

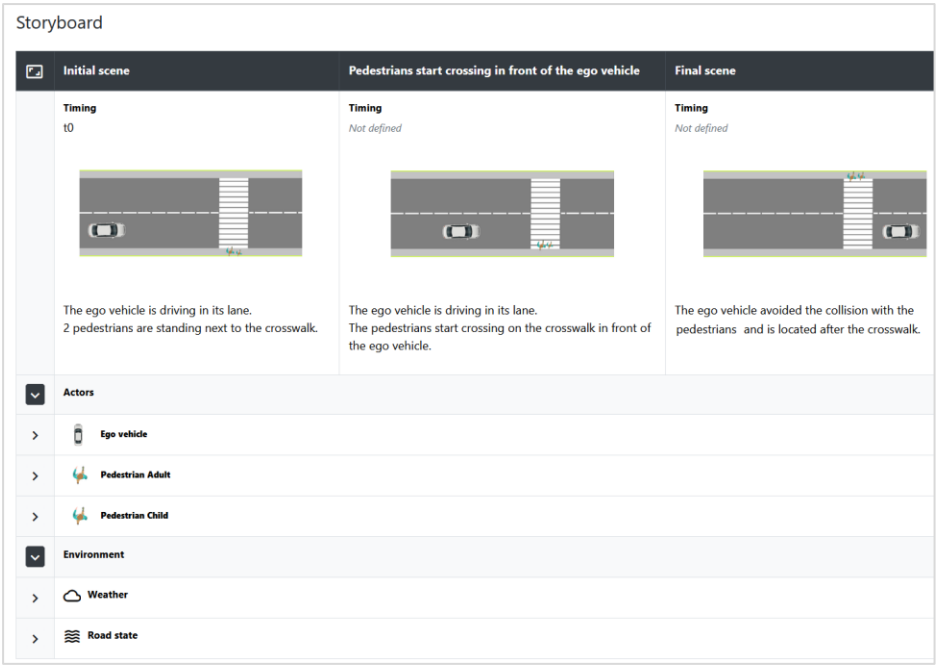


Figure 4. Storyboard of a logical scenario in MOSAR Scenario Manager: 2 pedestrians crossing on a crosswalk in front of ego (AI4CCAM Project)

The scenario modeled in the previous figures shows an example for a specific use case and structures the situation as well as the conditions that are to be described to perform simulation campaigns. The details of sampling of parameter ranges and therefore related coverage are not in the scope of this work.

In this context, trustworthiness criteria are considered, analyzed and applied before and after the scenarios are modeled.

Before the scenarios in the sense that the criteria analysis in the design phase yields as a result the parameters and ranges that should be considered in the scenarios, e.g. representativeness on actors in the scenarios, dynamic parameters for such actors, and environmental conditions considered as adverse to the system, among others.

Other criteria are also considered, analyzed and applied after synthetic data is obtained from the prescribed scenarios to simulation. *After* simulation test campaigns are performed, trustworthiness criteria related to the produced datasets can be assessed.

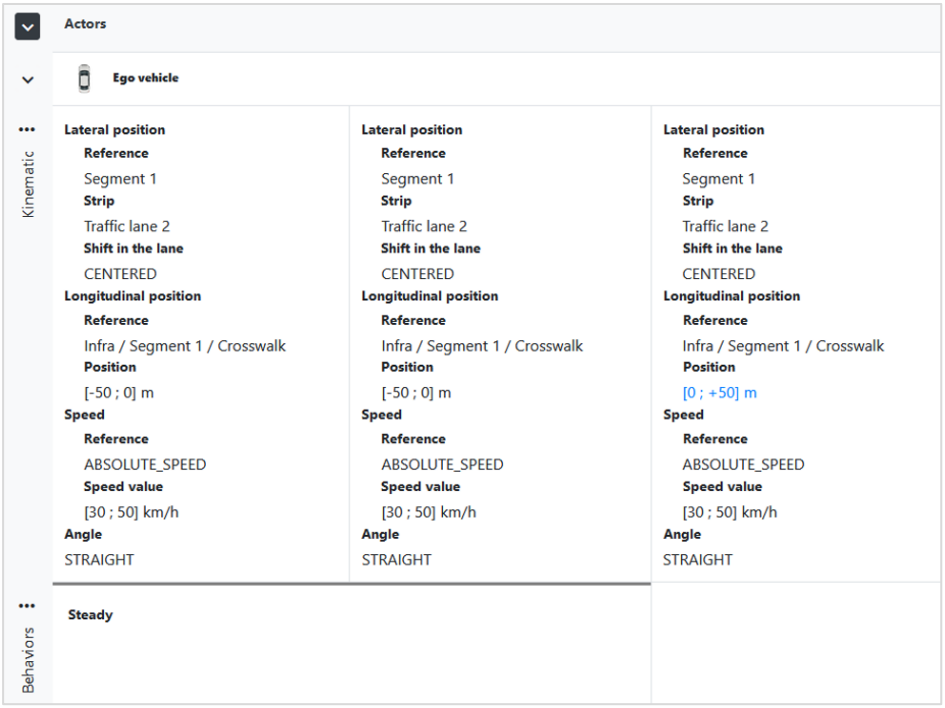


Figure 5. Ego vehicle parameters in MOSAR Scenario Manager (AI4CCAM Project)

5. Conclusions and Perspectives

This work depicts a first proposal for a method joining 3 fundamental currents to ensure trustworthy AI in CCAM systems. These are: 1) trustworthy requirements and criteria from a general standpoint, 2) trustworthiness assessment in all phases of a development pipeline (i.e. from design to evaluation and validation), and 3) Scenario-based testing for AVs as a common practice in this industry. A first attempt is presented to instantiating the approach for a case study namely using simulation. The method is considered as meet-in-the-middle since it suggests narrowing general trustworthiness-related attributes down to those applicable to the system under test and its operational context; while in parallel making use of largely used and well-established modeling and testing techniques such as scenario-based testing.

The proposed method is oriented towards AV-related applications, yet with proper adaptation it can be applied to other industrial sectors where proven good practices are already in place and cannot (and should not) be completely challenged or re-structured due to the integration of AI (ML based)-components. This work speaks to the proper synergies that should be built in order to integrate ML-based components in sound, robust and ongoing industrial processes in a trustworthy framework from the beginning of the process.

The following perspectives have been identified on different angles:

- On the application of the methodology: for the purposes and temporality of the work herein, this first pipeline of the Confiance.ai program proved fit for the immediate needs. As the program continues new results have emerged and its *end-to-end methodology* is now open to the public through what is referred to as the Confiance.ai body of knowledge, see [13]. This methodology revisits conventional and consolidated engineering pipelines and development cycles in order to include a more formal and structured approach to designing, specifying, developing, integrating and validating AI-based systems that can prove to be trustworthy by design. Deploying this end-to-end methodology in the CCAM ecosystem has not been done yet since it has been recently released and should prove useful as it can improve rigor and traceability through a thorough analysis of each step in the development cycle.

- On the scope of other research initiatives after Confiance.ai: similar initiatives are being pursued internationally. Some close examples include: Confiance.*ia* [7] the Quebecois program led by the Computer Research Institute of Montreal – CRIM, the project Zertifizierte Ki in Germany [14], and Responsible AI (RAI) UK [15]. The results of these programs should prove complementary to those obtained to this date.

- On the scope of the AI4CCAM project: next steps include simulation campaigns to build datasets that will be used to evaluate AI-models for VRU-trajectory prediction. Simultaneously, virtual reality test campaigns are being designed to get the VRUs perspective on a set of scenarios and assess the acceptance subjectively and through analysis of physiological parameters. A challenge remains on the representativeness of simulation vs real world testing and the need for robust, reliable simulation tools, frameworks, and comparison metrics and analyses.

Acknowledgments

The authors thank all the partners in the AI4CCAM project for their various contributions in the process of building the methodology, namely: Skoda (Czech Republic), Simula Research Laboratory (Norway), INLECOM Innovation (Greece), Barcelona Supercomputing Center (Spain), IMT and IMT Transfert (France), TTS Italia (Italy), Akkodis (France), UITP (Belgium), BVA (France), Deep Safety (Germany), Virtual Vehicle Research (Austria), and CNRS (France).

This work is funded by the European Commission through the AI4CCAM project (Trustworthy AI for Connected, Cooperative Automated Mobility) under grant agreement No 101076911.

References

[1] High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI," 2019. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[2] D. Fernandez Llorca and E. Gomez Gutierrez, "Trustworthy Autonomous Vehicles," Publications Office of the European Union, 2021.

[3] PEGASUS, J. Mazzega, D. Lipinski, U. Eberle, H. Schittenhelm and W. Wachenfeld, "PEGASUS Method - An Overview," 2019.

[4] C. Neurohr, L. Westhofen, T. Henning, T. de Graaff, E. Möhlmann and E. Böde, "Fundamental Considerations around Scenario-Based Testing," *CoRR*, 2020.

[5] P. Weissensteiner, G. Stettinger, S. Khastgir and D. Watzenig, "Operational Design Domain-Driven Coverage for the Safety Argumentation of Automated Vehicles," *IEEE Access*, vol. 11, pp. 12263-12284, 2023.

[6] O. Freiman, "Making Sense of the Conceptual Nonsense 'Trustworthy AI," *AI and Ethics*, vol. 4, 2022.

[7] Confiance IA Program, "Confiance IA," [Online]. Available: https://www.confianceia.ca/. [Accessed February 2024].

[8] High-Level Expert Group on Artificial Intelligence, "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment," 2020.

[9] AI4CCAM Project, "Trustworthy AI for Connected, Cooperative and Auromated Mobility," [Online]. Available: https://www.ai4ccam.eu/. [Accessed February 2024].

[10] A. Awadid, K. Amokrane-Ferka, H. Sohier, J. Mattioli and F. Adjed, "AI Systems Trustworthiness Assessment: State of the Art," in *Workshop on Model-based System Engineering and Artificial Intelligence - MBSE-AI Integration 2024*, Rome, Feb, 2024.

[11] Future of Life Institute, "High-level summary of the AI Act," 2024. [Online]. Available: https://artificialintelligenceact.eu/high-level-summary/.

[12] The Confiance.ai program, "Towards the engineering of trustworthy AI applications for critical systems," 2022.

[13] Confiance.ai Program, "Body of Knowledge, Confiance.ai," 2024. [Online]. Available: https://bok.confiance.ai/. [Accessed April 2024].

[14] Zertifizierte Ki Project, "Zertifizierte Ki," [Online]. Available: https://www.zertifizierte-ki.de/. [Accessed March 2024].

[15] Responsible AI UK, "Responsible AI UK," [Online]. Available: https://www.rai.ac.uk/. [Accessed February 2024].