







Design and Implementation of a Data Model for AI Trustworthiness Assessment in CCAM

Ruben Naranjo^{1,2}^a, Nerea Aranjuelo¹^b, Marcos Nieto¹^c, Itziar Urbieta^{1,2}^d,
Javier Fernández³^e and Itsaso Rodríguez-Moreno²^f

¹*Vicomtech Foundation, Basque Research and Technology Alliance (BRTA),
Mikeletegi 57, 20009, Donostia-San Sebastián, Spain*

²*University of the Basque Country (UPV/EHU), Donostia-San Sebastián, Spain*

³*Ikerlan Technological Research Center, Basque Research and Technology Alliance (BRTA),
José María Arizmendiarreta 2, 20500, Arrasate/Mondragón, Spain*

{*rmaranjo, naranjuelo, mnieto, turbieta*}@vicomtech.org, *javierfernandez@ikerlan.es, itsaso.rodriguez@ehu.es*

Keywords: AI Trustworthiness, Ontology, Data Model, CCAM, Trustworthiness Assessment.

Abstract: Amidst the growing landscape of trustworthiness-related initiatives and works both in the academic community and from official EU groups, there is a lack of coordination in the nature of the concepts used in these works and their relationships. This lack of coordination generates confusion and hinders the advances in trustworthy AI systems. This confusion is particularly grave in the CCAM domain given nearly all functionalities related to vehicles are safety-critical applications and need to be perceived as trustworthy in order for them to become available to the general public. In this paper, we propose the use of a defined set of terms and their definitions, carefully selected from the existing reports, regulations, and academic papers; and construct an ontology-based data model that can assist any user in the comprehension of those terms and their relationship to one another. In addition, we implement this data model as a tool that guides users on the self-assessment of the trustworthiness of an AI system. We use a graph database that allows making queries and automating the assessment of any particular AI system. We demonstrate the latter with a practical use case that makes an automated trustworthiness assessment based on user-inputted data.


1 INTRODUCTION


In the ever-evolving field of transportation, the incorporation of Artificial Intelligence (AI) has emerged as a crucial factor, especially in the domain of Connected Cooperative and Automated Mobility (CCAM). As communities embrace the revolutionary possibilities of smart transportation systems, the combination of connectivity, cooperation, and automation has ushered a new era characterised by improved effectiveness, safety, and convenience (Guerreiro Augusto et al., 2024). CCAM represents a paradigm shift in how society perceives and interacts with transportation. With vehicles seamlessly communicating with


each other and their surroundings, and automated features taking centre stage, the new possible mobility solutions are vast. From improved traffic management and reduced congestion to improved road safety and increased accessibility, the impact of AI on CCAM is far-reaching (Alonso Raposo et al., 2018).


However, amidst this rapid progress, the ethical dimensions and the need for trustworthy AI in CCAM cannot be underestimated. As vehicles become more interconnected and reliant on intelligent decision-making, ensuring AI systems' robustness, transparency, and ethical use becomes paramount (Kanak et al., 2022). Trustworthy AI not only safeguards the well-being of individuals on the road but also fosters public confidence in adopting these technologies. Trustworthiness plays an important role in almost any situation facilitated by Information Systems (IS), especially when uncertainty or potential undesirable outcomes are possible (Mcknight et al., 2011).


Various regulations and reports are emerging around this need (Commission, 2021; Fernan-


^a <https://orcid.org/0009-0002-3924-0194>

^b <https://orcid.org/0000-0002-7853-6708>

^c <https://orcid.org/0000-0001-9879-0992>

^d <https://orcid.org/0000-0001-7983-3651>

^e <https://orcid.org/0000-0002-4867-8115>

^f <https://orcid.org/0000-0001-8471-9765>

dez Llorca and Gomez Gutierrez, 2021; ISO, 2023), alongside a rapidly growing body of literature (Miller, 2019; Molnar, 2020; Langer et al., 2021; Graziani et al., 2023). However, these initiatives often lack coordination, leading to a fragmented approach to trustworthy AI. The primary hurdle to trustworthiness is establishing a concise definition of the desired goal. A significant challenge lies in the prevalence of fuzzy terms and overlapping definitions surrounding the concept of trustworthiness in different domains (Graziani et al., 2023). As the integration of AI in transportation advances, various stakeholders -such as researchers, policymakers, and industry practitioners- employ different terminologies to describe and assess the trustworthiness of AI systems tailored to their needs and perspectives. Along with this, AI is an area in increasing development; new subareas, terms, and definitions arise naturally and may cause conflicts with previous ones. This lack of standardized language and clear definitions can lead to ambiguity, hindering the establishment of universally accepted benchmarks for trustworthy AI and slowing down the progress in a common direction. In this paper, we propose to pave the way towards a standardised terminology based on existing regulations and literature from various perspectives (e.g., technical, legal). Rather than introducing new definitions, we focus on connecting existing ones and defining their relationships in the landscape of trustworthy AI. To achieve this, we present an ontology-based data model that integrates these terms, outlining their interrelations and requirements. We implement our data model in a graph database, creating a practical tool that allows users to evaluate the trustworthiness of AI systems by inputting specific information and properties. This tool provides a comprehensive overview of AI properties in relation to trustworthiness, based on existing initiatives, and regulations.

The main contributions of this paper are:

- We present an ontology that collects and organizes the most relevant terms related to AI trustworthiness, along with their relationships, in a controlled vocabulary based on current regulations, reports, and state-of-the-art of AI trustworthiness.
- We propose a data model to assess the trustworthiness of an AI system for CCAM based on a specific set of features provided by the interested user.
- We offer a tool that guides users through the self-assessment of the trustworthiness of an AI system.
- We demonstrate the application of our tool with a practical use case, assessing the trustworthiness of an AI-based pedestrian detection system.

- We make our data model publicly available to serve as a reference and support future research and development in this field¹.

2 RELATED WORK

In recent years, there has been an increasingly large number of publications and regulation initiatives concerning trustworthy AI. The European Government has been at the forefront of developing a regulatory framework for AI that is human-centric and trustworthy. The EU AI Act (Commission, 2021) is the first legal framework that sets global standards for AI systems. It defines four levels of risk for AI applications: minimal, limited, high, and unacceptable. Different regulations and obligations apply to each risk category. Minimal risk does not imply any obligation, limited risk conveys transparency obligations, high risk involves more specific requirements, and unacceptable risk leads to prohibited AI systems. The groundwork for AI regulation was laid with the publication of the Ethics Guidelines for Trustworthy AI by the High-Level Expert Group on Artificial Intelligence (AI HLEG) (Hleg, 2019). This established legality, ethical adherence, and robustness as the pillars of trustworthiness. These three pillars apply to the actors and processes involved in the AI systems (including their development, deployment, and use).

Regarding mobility, the European Joint Research Centre (JRC) published a report on Trustworthy Autonomous Vehicles (Fernandez Llorca and Gomez Gutierrez, 2021) as an application of the ideas of the AI HLEG to the CCAM sector (high-risk applications in the EU AI Act). The project reviews methods and tools to assess the safety, reliability, and explainability of Autonomous Vehicles (AVs) in real-world scenarios based on the different SAE levels of driving automation. It also provides policy recommendations and best practices for deploying AVs in the EU, taking into account ethical, legal, and social aspects. This report, along with other European initiatives (Hennemann et al., 2024), contributes to the EU's vision of creating a sustainable, intelligent, and safe mobility system for all. As can be noticed, the existing initiatives and regulations expand the dimensions of trustworthiness beyond traditional perspectives that primarily focus on safety to include new aspects such as fairness or privacy. Despite these expanded dimensions, safety remains a critical factor.

In the CCAM sector, the use of AI plays a crucial role in performing advanced autonomous vehi-

¹<https://github.com/rnarajo-vicomtech/ccam-tai-ontology>

cle functionalities, such as visual perception (Perez-Cerrolaza et al., 2023). These functionalities are often implemented in safety-related systems, and it is essential to provide evidence that errors or even the absence of errors do not lead to system malfunction. This is achieved by adhering to functional safety standards, such as those outlined by (IEC, 2010) for industrial domains and (ISO, 2018), for automotive domains. Consequently, functional safety is one of the aspects of trustworthiness that must be guaranteed.

However, traditional safety standards were not originally designed to accommodate AI in safety-related systems due to their development process, which relies on probabilistic models generated from training data. This is unlike traditional software components, which are coded from specifications. Significant efforts have been dedicated to the adoption of AI in safety-related systems in recent years (European Union Aviation Safety Agency (EASA), 2023; Hawkins et al., 2021), application rules (GmbH, 2020), and standards (VDA, 2023; IEC, 2023; IEC, TBD; ISO/IEC, 2024). Additionally, further standards are currently under development, such as (ISO, 2023), which will focus on safety and risk factors associated with AI within the road vehicle context.

However, since these reports were written by policymakers, law experts, and regulators, there are discrepancies between these reports and academic research on AI systems, such as the use of language and the end goal of trustworthiness (Gyevnar et al., 2023).

As a matter of fact, there is significant debate and disagreement within the academic community regarding the meanings of certain concepts. For instance, some authors use the terms interpretability and explainability interchangeably (Miller, 2019), while others give them distinct meanings (Molnar, 2020). Regarding the interrelations between the terms, we also find discrepancies. For example, some studies propose AI explainability as a means to support safety (Jia et al., 2022; Neto et al., 2022), viewing explainability as a requirement for achieving safety. However, this contrasts with European proposals that treat explainability and safety as distinct and independent concepts (Fernandez Llorca and Gomez Gutierrez, 2021; Hleg, 2019). This term definition and relationship disparity creates confusion and requires careful attention when reviewing research on these topics.

3 DATA MODEL FOR TRUSTWORTHY AI IN CCAM

We adopt a three-step methodology to craft our proposed data model for trustworthy AI within the con-

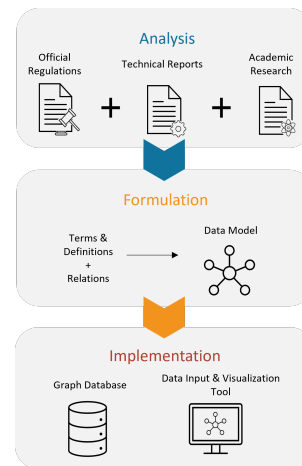


Figure 1: Data Model construction methodology.

text of CCAM. An overview of this methodology can be seen in Figure 1.

First, we conduct a comprehensive analysis of current regulations, official reports, and literature related to trustworthy AI for different stakeholders and CCAM. This aims to identify miscellaneous yet relevant terms and how these terms interact (Section 3.1).

Leveraging insights from our analysis, we develop an ontology to collect all relevant terms within their definitions and create a data model with the terms' interrelations and requirements (Section 3.2)

The implementation phase involves translating the data model into a tangible tool that can be used and consulted by a user for analyzing the trustworthiness of an AI system. We use a graph database and a visualization platform for this purpose (Section 3.3).

3.1 Understanding Trustworthiness

The initial step to understanding trustworthy AI is the definition of the terminology that forms its foundation. By delineating these terms, we can begin to understand their relationships that will later be used as a baseline for constructing our data model.

Regarding our proposed terms, concepts, and definitions, we have decided to prioritize official EU works, as future regulations and laws will be based on them. Our second priority has been the existing regulations concerning trustworthiness. We also consider academic research as a third step to address the gaps left by other works and to obtain an aligned proposal of all previous domains. The definitions have been selected following these criteria: prevalence of terms in the literature, closeness to the CCAM domain, and alignment with its meaning outside the AI context.

3.1.1 AI System and Properties of an AI System

According to the Council of the Organisation for Economic Co-operation and Development (OECD), an AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy (OECD, 2019). For example, an increasingly common ADAS function like Parking Assistance Systems (Naranjo et al., 2024) is usually considered an AI system because it needs to make decisions on whether something around the car is a parking slot, whether it is occupied, and whether it is suitable for parking, for which the use of machine learning models is frequent.

Based on the proposals and works analysed we have identified the following relevant properties for AI systems' trustworthiness:

- **Trustworthy:** AI system that is beneficial, and is perceived as beneficial, for the common good of humanity and the good for the immediate user(s), improving human welfare and freedom (Fernandez Llorca and Gomez Gutierrez, 2021). In CCAM environment, it is to note that a trustworthy AI system is considered trustworthy when inside its Operational Design Domain which should be stated to all stakeholders and depends on the level of vehicle automation (Committee, 2021).
- **Lawful:** Compliant with all applicable laws and regulations (Hleg, 2019). This includes laws and regulations on AI systems as well as laws and regulations affecting vehicles.
- **Ethical:** AI system that adheres to ethical principles and values. It aims to benefit, empower, and protect both individual human flourishing and the common good of society (Hleg, 2019).
- **Robust:** AI system that reliably behaves as intended while minimising unintentional and unexpected harm and preventing unacceptable harm to living beings or the environment. This also applies to potential changes in its operating environment or the presence of other agents that may interact with the system in an adversarial manner (Fernandez Llorca and Gomez Gutierrez, 2021).
- **Fair:** Assurance that individuals and groups are free from bias, discrimination and stigmatisation (Fernandez Llorca and Gomez Gutierrez, 2021).
- **Transparent:** AI system whose features, components, and procedures are open for external inspection by a stakeholder. This includes data, models, algorithms, training methods, decision mechanisms, and responsibility (ISO/IEC, 2020).
- **Safe:** AI system that does not lead to a state in which human life, health, property, or the environment is endangered (ISO/IEC, 2022).
- **Secure:** AI system that has resistance to intentional, unauthorized acts designed to cause harm or damage to a system (ISO/IEC, 2022). This includes acts performed physically (Pham and Xiong, 2021) or through any communication route such as Vehicle-to-Everything (V2X) communications (Ghosal and Conti, 2020).
- **Accurate:** AI system whose results of observations, computations, or estimates are close enough to the true values or the values accepted as being true (ISO/IEC, 2022). Accuracy is measured with different metrics depending on the target AI system and AV function. For example, from the perspective of overall vehicle performance (e.g., distance travelled) or from the technology layer perspective (e.g., precision and recall in object detection for scene understanding) (Fernandez Llorca and Gomez Gutierrez, 2021).
- **Reproducible:** AI system whose behaviour and results are consistent when repeated under the same conditions. (Hleg, 2019).
- **Explainable:** AI system that produces a clear, relevant, and accurate presentation of the reasoning, functioning, and behaviour behind its output, using partially or fully automated methods and directed to clearly defined stakeholders (Gyevnar et al., 2023). For instance, a route planning system should be able to explain the reason behind the route selection. Beyond this, a direct user of an autonomous driving system, must understand the state and driving capabilities of the system (Fernandez Llorca and Gomez Gutierrez, 2021), which might require additional efforts in terms of Human-Machine Interface (HMI) design.
- **Auditable:** AI system that enables the assessment of algorithms, data, and the design process (Fernandez Llorca and Gomez Gutierrez, 2021).
- **Accounted For:** AI system that has an accountable stakeholder for each step in its life cycle and functioning. This Stakeholder is answerable for actions, decisions and performance of the AI system (ISO/IEC, 2022).
- **Traceable:** AI system whose work items and artefacts are uniquely identifiable and can be tracked to the life-cycle step in which they were created (Fernandez Llorca and Gomez Gutierrez, 2021).

3.1.2 Stakeholders

Based on the ISO/IEC TS 5723 (ISO/IEC, 2022), a stakeholder is any individual, group, or organisation that can affect, be affected by, or perceive itself to be affected by a decision or activity. The perceived trustworthiness of an AI system might vary according to each stakeholder's needs and interests. We propose the following unified subcategories based on (Langer et al., 2021; Bhatt et al., 2020; Kanak et al., 2022).

- **Regulator:** Physical or legal entity that designs and publishes a regulation affecting an AI system.
- **Deployer:** Physical or legal entity that makes an AI system available to End-users. Different deployer profiles include original equipment manufacturers (OEMs), Tier 1 providers, Tier 2 providers, Tier 0.5 providers, or intermediaries.
- **Developer:** Physical or legal entity that designs, builds or maintains an AI system in any of its life-cycle steps, from sourcing of data to maintaining or enhancing of the system once it is deployed.
- **Domain Expert:** Physical or legal entity with extended knowledge of the domain of a system that ensures an AI system complies with one or several regulations, standards, or laws.
- **End-User:**
 - **Direct:** Physical or legal entity that makes use of an AI System (e.g. vehicle users).
 - **Indirect:** Physical or legal entity that is affected by an AI System in its functioning cycle. (e.g. external road users).

3.1.3 Related Terms

In addition to defining the properties of an AI system and the stakeholders affected, we identify relationships between terms and concepts. To allow to properly build those connections, we need to account for some additional terms that populate the trustworthiness landscape but are not directly involved with the AI system or the stakeholders:

- **Monitoring Process:** Continuous process or mechanism that tracks provenance, creation, and evolution of work items and artefacts in the life-cycle of an AI system, such as data collections, AI model versions, stakeholders involved, etc.
- **KPIs:** Key Performance Indicators (KPIs) in the context of an AI system refer to the quantitative results and metrics of the AI system.
- **Accuracy Requirements:** Thresholds that measure the accuracy of an AI system functions when

compared to KPIs. These are usually established by regulations, academic research or local laws.

- **Audition Process:** Process in which a domain expert reviews and tests the AI system against one or more regulations and, if successful, emits a certificate of compliance.
- **Regulation:** Binding legislative act of general application (Council of European Union, 2016).
- **Account-Giving Mechanism:** Automatic mechanism that checks the life-cycle and functioning status of the AI system and assigns accountability to the appropriate stakeholder.
- **Explanation:** Clear, relevant, and accurate presentation of the reasoning, functioning, and behaviour behind the output of an AI system.
- **Explainability Mechanism:** Mechanism or process that generates one or more explanations for the output of an AI system. There are several emerging taxonomies regarding explainability; given the lack of standardization, we have chosen a high-level categorization that organizes explainability mechanisms depending on strategy.
 - **Explaining Method:** Most commonly known post-hoc explanations. These methods usually generate explanations to interpret black-box models (Hassija et al., 2024). Explanations can be local (for single input values, e.g. saliency maps) or global (which features contribute more to the overall model).
 - **Interpretable Design:** This is the kind of mechanism that comes implicitly in interpretable or white-box models. The functioning of this kind of model is designed to be clear so that the explanation of the output is the model itself. e.g. classification trees.
 - **Evaluation Process:** Strategy that aims to generate explanations based on exhaustive testing of the system in a very extensive set of data with the objective of predicting the system behaviour under different circumstances. This kind of process is very common in CCAM applications where auto-makers test functionalities in hundreds or thousands of scenarios.

3.2 Data Model Design

An ontology allows organising the terms related to a domain, trustworthiness in this case, into a controlled vocabulary and capturing the relationships between objects or concepts and the properties that can be used to describe them. We create an ontology as the base of our data model.

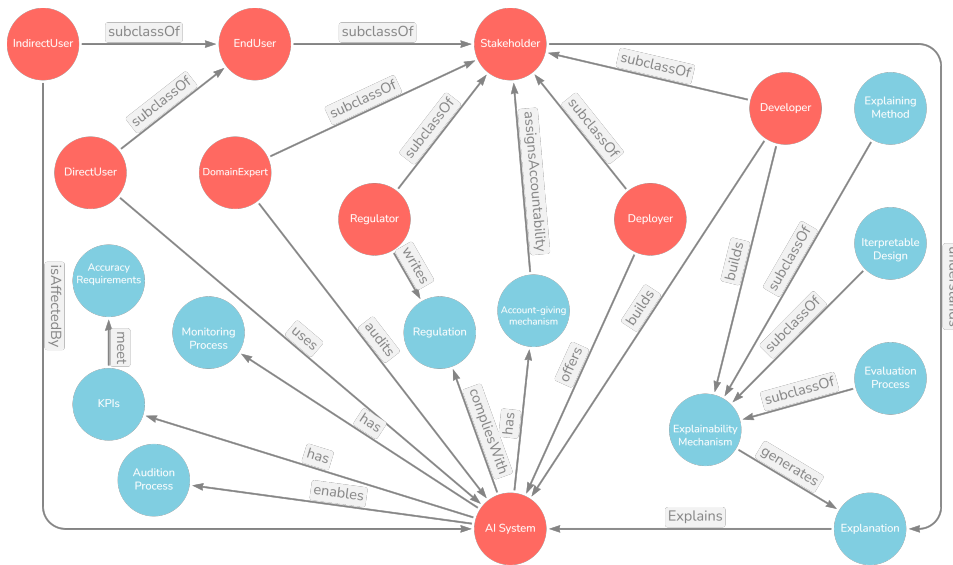


Figure 2: Complete data model. The main classes are in red shades, and the supporting classes are in blue shades.

3.2.1 Elements of the Data Model

We propose AI System and Stakeholder to be the two main classes acting as pillars for building our data model. Since the data model’s purpose is to analyze an AI system’s trustworthiness, the AI system itself is one of the main building blocks of the data model. The AI system class has several features or properties corresponding to some of the terms described in Section 3.1 and the risk level defined in the EU AI Act. These features can be seen in Table 1. It is crucial to note that although the defined terms (e.g., explainable, transparent) are general when applied as properties of the AI System class within the data model, the notation “<term>To” is employed in certain instances to specify that the term should consistently relate to a particular Stakeholder subclass or a group of Stakeholder subclasses (e.g., ExplainableTo[Developer], TransparentTo[Direct-user, Deployer]).

The second pillar, Stakeholder, gains importance when analyzing the relationships between the terms described in Section 3.1. As some of these terms, e.g. explainability, are evaluated by different stakeholders, the Stakeholder itself and its subclasses should be a key component in analyzing the trustworthiness of the AI system.

In addition to these pillars, we propose a set of supporting classes on our data model, that will help determine the trustworthiness of an AI system. These supporting classes (e.g., regulation, explainability mechanism) enable the analysis of the AI system and its features through the use of some predefined rules. Consequently, these classes are required to obtain the properties of the AI system. The whole

Table 1: Properties of the AI System and their data types.

Property	Data Type	Property	Data Type
Lawful	bool	EthicalTo	array[string]
Traceable	bool	Deterministic	bool
ExplainableTo	array[string]	Reproducible	bool
Auditable	bool	Accurate	bool
AccountedFor	bool	Secure	bool
TransparentTo	array[string]	Robust	bool
Fair	bool	TrustworthyTo	array[string]
Safe	bool	RiskLevel	integer

set of classes including the two pillars and the supporting classes can be seen in Figure 2.

3.2.2 Rules for Trustworthiness Assessment

Once the data model is defined, we design a set of rules that assign features to the AI System class based on user-provided input. These rules are used to evaluate the trustworthiness of the AI system defined by the user.

These rules are the result of our analysis of the relationships of the terms in Section 3.1, and are based on proposals from the literature, but leave some tasks that the user shall previously solve. Since this data model encompasses complex high-level terms such as fairness, safety, or regulation compliance, the automatic analysis of these elements is out of the scope of this work. Especially given that the analysis of many of these concepts is still an open question that is continually being researched (Parraga et al., 2023; Caviglione et al., 2023; Buyl and De Bie, 2024). The advances and research in these domains will complement the work presented in this paper.

The rules are presented summarised in Figure 3 and as pseudo-code in the Appendix. For example, for a given stakeholder, if the AI system is lawful,

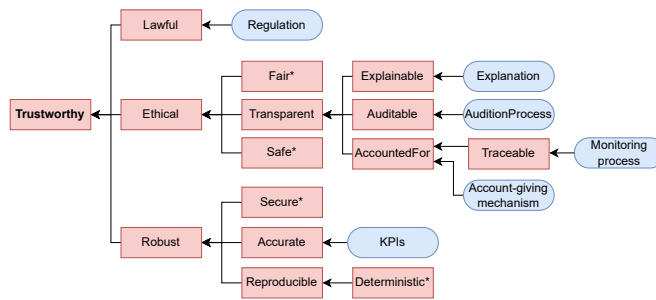


Figure 3: Summary of rules for the data model. Each property (in red) entails compliance with other properties or auxiliary classes of the data model (in blue). Asterisk marks user-inputted data. Note that explainable is for a given stakeholder, and consequently transparent, ethical, and trustworthy.

ethical, and robust, then the system is trustworthy.

and request data. Internally, the API uses *SPARQL* to query the ontology within *GraphDB*.

3.3 Data Model Implementation

To bring the data model into practical use, a multi-step process has been designed to ensure usability and scalability. This process includes constructing an ontology, loading it into a graph database, and enabling user interaction through a REST API.

- **Ontology Construction:** the foundation of the data model implementation begins with constructing the ontology using *Protégé*², an open-source ontology editor. The classes, properties, data types, and relationships in the ontology are defined according to the detailed definitions provided in Section 3. This tool allows to meticulously define and organize the various classes, properties, and relationships that form the data model. These elements are defined according to the detailed definitions provided in Section 3 of this paper, which are derived from our extensive analysis of current regulations and literature.
- **Loading the Ontology into the semantic graph database:** once the ontology is constructed, it is loaded into *GraphDB*³, a high-performance graph database optimized for handling ontologies and linked data. We chose *GraphDB* for its capability to store complex relationships and support *SPARQL*, a powerful query language used for retrieving and manipulating data in RDF format.
- **REST API development:** To facilitate user interaction with the ontology and database, a REST API is developed using *Flask*⁴, a lightweight web framework for *Python*. The REST API acts as an intermediary, enabling users to seamlessly enter

4 PRACTICAL APPLICATION OF THE DATA MODEL TO ASSESS TRUSTWORTHINESS

This section outlines the tool designed with user-friendliness in mind, developed as part of the use case demonstrations in this work. It is decomposed into two subsections, the first one outlines the interface of the tool and the later the practical application that has been evaluated in this paper as an example.

Figure 4: Input panel of the assessment tool.

4.1 Tool Interface

Using the database implementation described in Section 3.3, we build a graphing and data input tool on

²<https://protege.stanford.edu/>

³<https://graphdb.ontotext.com/documentation/10.7/>

⁴<https://flask.palletsprojects.com/en/3.0.x/>

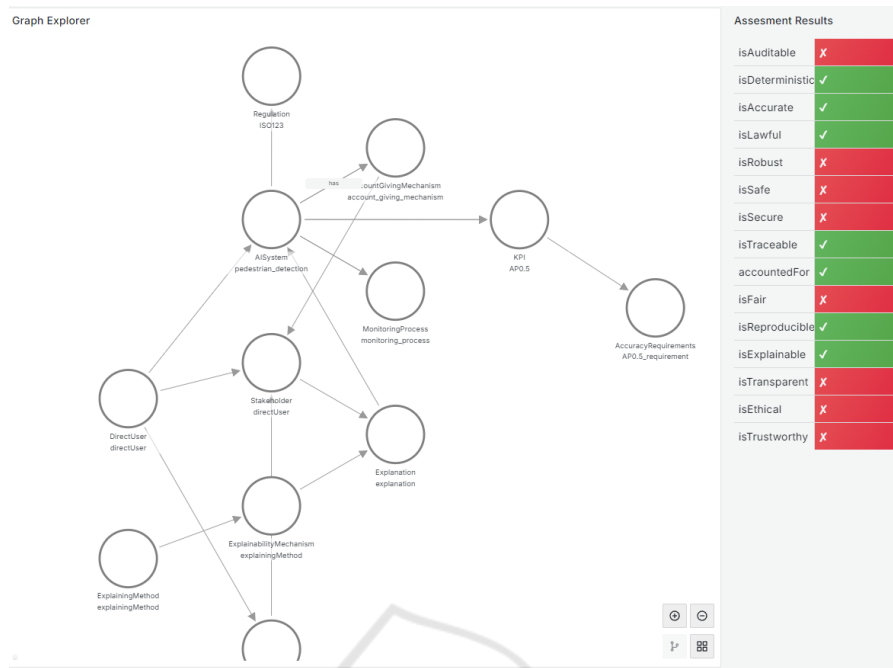


Figure 5: Graph visualization and assessment result panels of the assessment tool. Please zoom-in for better visualization.

*Grafana*⁵. Grafana is a powerful open-source analytics and monitoring tool extended with community-created plugins. We use some of these plugins to add useful functionality, such as data input with the *Data Manipulation Plugin*⁶ and REST API communication through the *Infinity Data Source Plugin*⁷.

With this setup, we construct a dashboard composed of three main panels:

1. The input data panel allows the user to enter the required information about their AI system. This panel uses the Data Manipulation Plugin to build an input form with the necessary data required from the user. When the submit button is clicked, it constructs a POST request and sends it to the REST API, which communicates with the database and updates its values.
2. The graph visualization panel displays a graphical representation of the inputted data and its relationships, following the structure of the data model. It uses the Infinity Data Source Plugin to perform GET requests to the REST API, which retrieves data from the database.
3. The assessment result panel shows the calculated properties of the AI system using a *Table Panel*

⁵<https://grafana.com/>

⁶<https://grafana.com/grafana/plugins/volkovlabs-form-panel/>

⁷<https://grafana.com/grafana/plugins/yesoreyeram-infinity-datasource/>

from Grafana. The panel uses the Infinity Data Source Plugin to perform a GET request to the REST API and displays the properties of the AI system as a list, showing which of them are fulfilled and which of them are not.

4.2 Practical Application

To demonstrate the use of the tool, we show a practical example in which we want to assess the trustworthiness of a pedestrian detection system we have developed using a Deep Neural Network (DNN).

The first step is to enter the required data in the input data panel (Figure 4). In this panel, the user shall select and enter data about the AI system to be assessed. These data are some of the high-level concepts described in Section 3.1, such as whether the system is fair, deterministic, or has a monitoring process. This information should be gathered or analyzed by the user in a prior step.

In our case, we know that our pedestrian detection model is deterministic and we have an account-giving and monitoring process for its entire life-cycle. We also know that it needs to meet some Average Precision (AP) requirements, and it does. The model has an explainability mechanism in place in the form of an explaining method, in this case, we have a variant of D-RISE (Petsiuk et al., 2021) that generates explanations of the output that we know our target stakeholder (direct users) understand. Since we did not manually

assess some aspects, such as auditability and security, we select false and/or unknown values in the corresponding columns in the input data panel.

Once the data has been collected, the graph visualization and assessment result panels will be populated. In them, we can explore the relationships between the data model classes and the properties of the trustworthiness landscape that have been calculated for the assessed AI system. These two panels are shown in Figure 5.

In the case of our pedestrian detection model, we can see that it fulfils some of the requirements such as being reproducible and explainable. However, it fails in other aspects such as transparency and robustness, requiring the use of additional mechanisms to achieve a trustworthy system.

5 CONCLUSIONS

In this paper, we presented an ontology that systematically collects and organizes the most relevant terms related to AI trustworthiness. Our ontology, based on current regulations, reports, and state-of-the-art academic research, provides a controlled vocabulary that can be universally adopted, enhancing clarity and consistency in discussions about AI trustworthiness.

Based on the proposed ontology, we built a data model that can help interested users to evaluate AI systems based on a well-defined set of features, ensuring a tailored and relevant assessment process of the trustworthiness.

We provide a practical tool that facilitates self-assessment of AI system trustworthiness. The applicability of this tool is demonstrated through a practical application in a common CCAM domain use case. This demonstration highlights the tool's practical utility and effectiveness in real-world scenarios, providing a concrete example of its application.

In summary, our contributions provide a structured and practical approach to assessing AI trustworthiness, supporting both theoretical understanding and practical implementation. The work presented in this paper contributes to the development of open-source tools, paving the way towards creating trustworthy AI systems, with a particular focus on the CCAM domain.

ACKNOWLEDGEMENTS

This work has received funding from the Basque Government under project AutoTrust of the program

ELKARTEK-2023, by the European Union's Horizon Programme under Grant 101076754 — AIthena Project and by Spanish Ministry of Science and Innovation under project PLEC2023-010240, CAPSUL-IA.

REFERENCES

- Alonso Raposo, M., Grosso, M., Després, J., Fernández Macías, E., Galassi, C., Krasenbrink, A., Krause, J., Levati, L., Mourtzouchou, A., Saveyn, B., et al. (2018). An analysis of possible socio-economic effects of a Cooperative, Connected and Automated Mobility (CCAM) in Europe. *European Union*.
- Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M., and Eckersley, P. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 648–657.
- Buyl, M. and De Bie, T. (2024). Inherent limitations of AI fairness. *Communications of the ACM*, 67(2):48–55.
- Caviglione, L., Comito, C., Guarascio, M., and Manco, E. (2023). Emerging challenges and perspectives in deep learning model security: A brief survey. *Systems and Soft Computing*, 5:200050.
- Commission, E. (2021). Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. *Eur Comm*, 106:1–108.
- Committee, O.-R. A. D. O. (2021). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles*. SAE International.
- Council of European Union (2016). Treaty on the functioning of the European Union. article 288. http://data.europa.eu/eli/treaty/tfeu_2016/art_288/oj.
- European Union Aviation Safety Agency (EASA) (2023). EASA Concept Paper: guidance for Level 1 & 2 machine learning applications.
- Fernandez Llorca, D. and Gomez Gutierrez, E. (2021). Trustworthy autonomous vehicles. Technical Report KJ-NA-30942-EN-N (online), Luxembourg (Luxembourg).
- Ghosal, A. and Conti, M. (2020). Security issues and challenges in v2x: A survey. *Computer Networks*, 169:107093.
- GmbH, V. V. (2020). Design and Trustworthiness of autonomous/cognitive systems.
- Graziani, M., Dutkiewicz, L., Calvaresi, D., Amorim, J. P., Yordanova, K., Vered, M., Nair, R., Abreu, P. H., Blanke, T., Pulignano, V., et al. (2023). A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. *Artificial intelligence review*, 56(4):3473–3504.
- Guerreiro Augusto, M., Acar, B., Soto, A. C., Sivrikaya, F., and Albayrak, S. (2024). Driving into the future: a cross-cutting analysis of distributed artificial intelligence, CCAM and the platform economy. *Autonomous Intelligent Systems*, 4(1):1–11.

- Gyevnar, B., Ferguson, N., and Schafer, B. (2023). Bridging the transparency gap: What can explainable AI learn from the AI Act? In *26th European Conference on Artificial Intelligence*, pages 964–971. IOS Press.
- Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., and Hussain, A. (2024). Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1):45–74.
- Hawkins, R., Paterson, C., Picardi, C., Jia, Y., Calinescu, R., and Habli, I. (2021). Guidance on the Assurance of Machine Learning in Autonomous Systems (AM-LAS).
- Hennemann, M., Ebner, G. K., Karsten, B., Lienemann, G., and Wienroeder, M. (2024). *Data Act: An Introduction*. Nomos Verlagsgesellschaft mbH & Co. KG.
- Hleg, A. (2019). Ethics guidelines for trustworthy AI. *B-1049 Brussels*.
- IEC (2010). IEC 61508(-1/7): Functional safety of electrical / electronic / programmable electronic safety-related systems.
- IEC (2023). ISO/IEC WD 5338 Information technology — Artificial intelligence — AI system life cycle processes.
- IEC (TBD). IEC TS 6254 - Information technology — Artificial intelligence — Objectives and approaches for explainability of ML models and AI systems (Under Development).
- ISO (2018). ISO 26262(-1/11) Road vehicles – Functional safety.
- ISO (2023). ISO/CD PAS 8800 Road Vehicles — Safety and artificial intelligence.
- ISO/IEC (2020). ISO/IEC TR 24028 - Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence.
- ISO/IEC (2022). ISO/IEC TS 5723 - Trustworthiness — Vocabulary.
- ISO/IEC (2024). ISO/IEC TR 5469 Artificial intelligence — Functional safety and AI systems.
- Jia, Y., McDermid, J., Lawton, T., and Habli, I. (2022). The role of explainability in assuring safety of machine learning in healthcare. *IEEE Transactions on Emerging Topics in Computing*, 10(4):1746–1760.
- Kanak, A., Ergün, S., Atalay, A. S., Persi, S., and Karcı, A. E. H. (2022). A review and strategic approach for the transition towards third-wave trustworthy and explainable ai in connected, cooperative and automated mobility (ccam). In *2022 27th Asia Pacific Conference on Communications (APCC)*, pages 108–113. IEEE.
- Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., and Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)?—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296:103473.
- Mcknight, D. H., Carter, M., Thatcher, J. B., and Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on management information systems (TMIS)*, 2(2):1–25.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.
- Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.
- Naranjo, R., Sintés, J., Pérez-Benito, C., Alonso, P., Delgado, G., Aranjuelo, N., and Jevtić, A. (2024). Park marking detection and tracking based on a vehicle on-board system of fisheye cameras. In *International Conference on Robotics, Computer Vision and Intelligent Systems*, pages 31–46. Springer.
- Neto, A. V. S., Camargo, J. B., Almeida, J. R., and Cugnasca, P. S. (2022). Safety assurance of artificial intelligence-based systems: A systematic literature review on the state of the art and guidelines for future work. *IEEE Access*, 10:130733–130770.
- OECD (2019). Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/044.
- Parraga, O., More, M. D., Oliveira, C. M., Gavenski, N. S., Kupssinskü, L. S., Medronha, A., Moura, L. V., Simões, G. S., and Barros, R. C. (2023). Fairness in deep learning: A survey on vision and language research. *ACM Computing Surveys*.
- Perez-Cerrolaza, J. et al. (2023). Artificial Intelligence for Safety-Critical Systems in Industrial and Transportation Domains: A Survey. *ACM Comput. Surv.*
- Petsiuk, V., Jain, R., Manjunatha, V., Morariu, V. I., Mehra, A., Ordonez, V., and Saenko, K. (2021). Black-box explanation of object detectors via saliency maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11443–11452.
- Pham, M. and Xiong, K. (2021). A survey on security attacks and defense techniques for connected and autonomous vehicles. *Computers & Security*, 109:102269.
- VDA (2023). Automotive SPICE® Process Assessment / Reference Model Version 4.0.

APPENDIX

Here we present the set of rules designed to make the self-assessment of an AI system using the proposed data model. The rules are presented as pseudo-code in Algorithms 1 to 11.

```

if AISystem.deterministic then
  | AISystem.reproducible = True
end

```

Algorithm 1: Reproducibility rule.

```

AISystem.accurate = True
foreach KPI in KPIS do
  | if not (KPI MEETS
  | AccuracyRequirements) then
  | | AISystem.accurate = False
  | end
end

```

Algorithm 2: Accuracy rule.

```

if AISystem.isSecure
and AISystem.accurate
and AISystem.reproducible then
  | AISystem.robust = True
end

```

Algorithm 3: Robustness rule.

```

AISystem.lawful = True
foreach Regulation in AffectingRegulations
do
  | if not AISystem COMPLIESWITH
  | Regulation then
  | | AISystem.lawful = False
  | end
end

```

Algorithm 4: Lawfulness rule.

```

if AISystem HAS MonitoringProcess then
  | AISystem.traceable = True
end

```

Algorithm 5: Traceability rule.

```

if AISystem HAS Account-givingMechanism
and AISystem.traceable then
  | AISystem.accountedFor = True
end

```

Algorithm 6: Accountability rule.

```

if AISystem ENABLES AuditionProcess then
  | AISystem.auditable = True
end

```

Algorithm 7: Auditability rule.

```

foreach Stakeholder in Stakeholders do
  | if AISystem HAS
  | ExplainabilityMechanism
  | and Stakeholder UNDERSTANDS
  | Explanation then
  | | AISystem.explainableTo.append(
  | | Stakeholder)
  | end
end

```

Algorithm 8: Explainability rule.

```

foreach Stakeholder in Stakeholders do
  | if AISystem.accountedFor
  | and AISystem.auditable
  | and
  | | AISystem.explainableTo[Stakeholder]
  | | then
  | | | AISystem.transparentTo.append(
  | | | Stakeholder)
  | | end
end

```

Algorithm 9: Transparency rule.

```

foreach Stakeholder in Stakeholders do
  | if AISystem.fair AND
  | | AISystem.transparentTo[Stakeholder]
  | | AND AISystem.safe then
  | | | AISystem.ethicalTo.append(
  | | | Stakeholder)
  | | end
end

```

Algorithm 10: Ethicalness rule.

```

foreach Stakeholder in Stakeholders do
  | if AISystem.lawful
  | and AISystem.ethicalTo[Stakeholder]
  | and AISystem.robust then
  | | AISystem.trustworthyTo.append(
  | | Stakeholder)
  | end
end

```

Algorithm 11: Trustworthiness rule.