SAE INTERNATIONAL

ARTICLE INFO
Article ID: 12-08-01-0010

doi:10.4271/12-08-01-0010

A Methodology to Enhance Transparency for Trustworthy Artificial Intelligence for Cooperative, Connected, and Automated Mobility

Paola Natalia Cañas,[1,2] Marcos Nieto,[1] Oihana Otaegui,[1] and Igor Rodriguez[2]

[1]Basque Research and Technology Alliance (BRTA), Fundación Vicomtech, Spain
[2]University of the Basque Country (UPV/EHU), Spain

Abstract

In this research, we propose a set of reporting documents to enhance transparency and trust in artificial intelligence (AI) systems for cooperative, connected, and automated mobility (CCAM) applications. By analyzing key documents on ethical guidelines and regulations in AI, such as the Assessment List for Trustworthy AI and the EU AI Act, we extracted considerations regarding transparency requirements. Recognizing the unique characteristics of each AI system and its application sector, we designed a model card tailored for CCAM applications. This was made considering the criteria for achieving trustworthy autonomous vehicles, exposed by the Joint Research Centre (JRC), and including information items that evidence the compliance of the AI system with these ethical aspects and that are also of interest to the different stakeholders. Additionally, we propose an MLOps Card to share information about the infrastructure and tools involved in creating and implementing the AI system.

History

Received: 02 Jan 2024
Revised: 11 Jun 2024
Accepted: 15 Aug 2024
e-Available: 02 Sep 2024

Keywords

Trustworthy AI, CCAM, Transparency, Explainability, Model cards, Data cards, MLOps

Citation

Cañas, P., Nieto, M., Otaegui, O., and Rodriguez, I., "A Methodology to Enhance Transparency for Trustworthy Artificial Intelligence for Cooperative, Connected, and Automated Mobility," *SAE Int. J. of CAV* 8(1):125–139, 2025, doi:10.4271/12-08-01-0010.

ISSN: 2574-0741
e-ISSN: 2574-075X

This article is part of a focus issue on the Safety of AI-Based Systems.

1. Introduction

In these changing times, the application of AI (artificial intelligence) algorithms has caused a paradigm shift across various sectors and everyday activities. This technological progress is perceived with great enthusiasm but also with great uncertainty. AI has earned its place in society by demonstrating its potential for analysis and, more importantly, for assisting humans. However, this has not come without concerns and questions about the functioning of these systems and their potential risks. Published reports already highlight situations when these systems fail [1, 2] or are misused by humans [3, 4], questioning the reliability of AI.

The cooperative, connected, and automated mobility (CCAM) sector has been potentiated due to advances in AI algorithms, with the development of assisted or automated driving functions, including SAE-L2-4 features [5] [e.g., advanced driver assistance systems (ADAS), occupant monitoring systems (OMS), automated parking systems (APS), etc.] to achieve safer and more efficient transport. Mobility is a service sector that humanity interacts daily with, meaning that the success of any application in this field will depend on its users' trust. This implies the necessity to develop trustworthy systems as a crucial prerequisite for integrating and accepting them into society. When such systems made use of AI technology, two main questions arise: What is trustworthy AI? How to make an AI trustworthy?

Work on the definition and research in this topic has begun with a bumpy start since there are many perspectives on what trust is due to the diversity of human cultures and beliefs. This discussion has produced many documents [6, 7, 8], research papers [9], and opinions [10] about how AI can be trusted and whether it can be trusted at all [11]. In Europe, there have been significant steps toward the definition of trustworthy AI with the publication of the Ethics Guidelines for Trustworthy AI in 2019 [12] by the High-Level Expert Group on Artificial Intelligence (AI HLEG). This document aims to guide the ethical use and regulation of AI, stating that these systems must not only be technically robust and reliable but also ethical. This concept is derived from the Fundamental Human Rights, orienting the development of these systems in a way that is compliant and respectful of human rights: an agreed basic norm of coexistence.

The AI HLEG proposes a list of seven key requirements (KR) for AI systems to be trustworthy: KR1 Human Agency and Oversight, KR2 Technical Robustness and Safety, KR3 Privacy and Data Governance, KR4 Transparency, KR5 Diversity, Non-discrimination and Fairness, KR6 Societal and Environmental Well-being, and KR7 Accountability. To make these principles more practical and applicable, they further elaborated a comprehensive list of specific and technical considerations that should be addressed during the development, deployment, testing, and monitoring of an AI system. This document is the Assessment List for Trustworthy AI (ALTAI) [13]. During the same year, the intergovernmental Organization for Economic Cooperation and Development (OECD) made public a set of AI principles [14] to become standard and provide a foundation for governments to legislate and regulate AI. These principles promote the creation and use of a trustworthy AI by (i) inclusive growth, sustainable development, and well-being, (ii) human-centered values and fairness, (iii) transparency and explainability, (iv) robustness, security, and safety, (v) accountability.

On a normative level, the European Union has published an AI Act in 2021 [15]. This document is the closest to achieving a comprehensive legal framework for AI in the European Union by the date of writing this article. The AI Act is aligned with the OECD principles, adopting its definition of AI. It classifies the systems according to the risks in their implementation: (i) unacceptable risk, (ii) high risk, (iii) limited risk, and (iv) low or minimal risk applications. Each risk category is subject to different regulations and obligations, and the applications will be regulated depending on the level of risk they fall under. Systems categorized as limited risk would be subject to a limited set of transparency obligations. High-risk systems are subject to more specific requirements. Unacceptable risks are banned, and low-risk applications have no obligations. The AI systems related to CCAM applications are considered high-risk [16].

According to the AI HLEG, given the context-specificity of AI systems, the implementation of these guidelines needs to be adapted to the particular application. Moreover, the necessity of an additional sectorial approach to complement the more general horizontal framework should be investigated [12]. This exploration in CCAM can be evidenced in [17] and especially in the Trustworthy Autonomous Vehicles—Assessment criteria for trustworthy AI in the autonomous driving domain study [16] by the Joint Research Centre (JRC). The JRC study offers a comprehensive overview of the steps needed to achieve trustworthiness at each of their proposed five key technology layers of autonomous vehicles (AVs): localization, dynamic scene understanding, path planning, control, and user interaction. The JRC study is a translation of the seven KR that AI systems should meet in order to be trustworthy from the ALTAI to the context of AVs. Regarding transparency, it primarily focuses on the technical perspective with no mention of specific communication details, documentation, or information reporting. Their main contribution to this article is the definition of a list of criteria with which the KR of the ALTAI can be accomplished in the AVs' domain.

Initial efforts have been made toward defining a peaceful and ethical framework for AI integration into society, but there is still a long way to go. This article builds upon the concept of trustworthy AI, proposing a methodology that aligns with the guidelines and obligations outlined in the ALTAI and AI Act proposal, adapted

to CCAM applications through the analysis of the requirements and assessment criteria of the JRC. This work contributes to materialized practices for addressing the transparency requirement, offering a methodology composed of three report documents: Data Cards, CCAM-tailored Model, and MLOps (Machine Learning Operations) Cards.

The article is organized as follows: (i) The introduction covers the importance of building trustworthy AI systems and reviewing key documents that tackle this matter. It also includes information about AI reporting and the different templates proposed to disclose information about machine learning (ML) models. (ii) In the methodology, the template of the Model Card is presented. Mapping each of its information items with an ethical requirement or obligation. The MLOps Card is also part of the section, along with an analysis of the stakeholders of these reporting documents. (iii) The discussion section explores the need for standardization, intellectual property concerns, and practical communication formats, drawing lessons from other industries. (iv) Finally, the conclusions section wraps up the contributions of this article.

Transparency

One major pillar for Trustworthy AI is transparency; both the ALTAI document and the AI Act emphasize the importance of transparency in AI systems to achieve trustworthiness and lawfulness. However, the definition of transparency and the means to accomplish it are not the same across the different documents of AI guidelines, ethics, and regulation. In [18], the study showed this discrepancy, with transparency being the most included ethical term in guidelines/regulatory documents. But, when defining it, some include other terms or means to achieve it, like explainability, explicability, understandability, interpretability, communication, disclosure, showing, and the like.

Transparency, as defined in the ALTAI, requires traceability, explainability, and open communication about the limitations of the AI system. In the AI Act, there is no precise definition of transparency; however, it mentions that, due to the difficulty in making the algorithms self-explanatory, high-risk AI systems should "*be accompanied by relevant documentation and instructions of use and include concise and clear information, including in relation to possible risks to fundamental rights and discrimination of the persons who may be affected by the system in light of its intended purpose, where appropriate*." Additionally, Article 11 (Technical Documentation) states that high-risk systems must have technical documentation prepared that includes the points presented in Annex IV. Article 13 (Transparency and provision of information to users) says these systems shall be "*accompanied by instruction for use that include concise, complete, correct, and clear information that is relevant, accessible and comprehensible to users*," also providing a list of information points to include.

Consequently, communication requirements to accomplish transparency involve reporting relevant information from the AI system [19]. This information should cover the entire lifecycle of the AI system: from preparation, development, testing, deployment, and implementation. Table 1 summarizes related regulations, reports, and recommendations that propose guidelines for building ethical and trustworthy AI; they include transparency as a key requirement. The table shows the provided definition of transparency, the ethical terms included when defining it or as a means to achieve transparency, and the details about what information should be reported. Although this article does not intend to provide a semantic or philosophical debate about the terms and the relation between them, causality, or composition, we review the requirements related to transparency, especially regarding communication and disclosure, to find their applicability in CCAM systems.

Reporting Documents

Despite the popularization of AI, there is no standard way to report relevant information about these systems. Now, with the transparency requirement, it has become mandatory to report information. However, questions remain about what information is relevant to report, who are the end users of these documents, what is the most effective format for presentation, and the like [18, 19]. Also, although principles and regulations, as outlined in Table 1, provide some direction on what to report, clear guidance remains largely broad and ambiguous.

Several proposals for AI-related reporting documents have already been published. In 2018, Mitchell et al. [20] introduced the concept of a Model Card: a document that discloses information about a ML model, such as its intended use, development, performance, and relevant evaluations under a variety of conditions and across different impacted human groups in the target application domain to demonstrate its fairness. This was the first documentation framework to be proposed and remains relevant today, with examples of its implementation found in [21]. More recently, the open-source model repository platform Hugging Face has launched a Model Card Creator App with a new and improved template based on some user studies on Model Card usage [22].

Another essential proposal is the Factsheets [23]. The authors inspired this document in the supplier's declarations of conformity, an already existing and standardized document that promotes transparency in other industries. It is essentially a collection of information about how an AI model or service was developed and deployed.

The Method Cards [24] is an approach targeting report consumers who are specifically other ML engineers. It allows these consumers or stakeholders to reproduce the models, understand the rationale behind their designs, and introduce adaptations in an informed way. The information comprises both prescriptive and

TABLE 1 Definition of transparency, means to achieve it, and recommend reporting information.

Title	Date	Definition	Means	Reporting info
ALTAI	2019	Closely linked with the principle of explicability and encompasses transparency of elements relevant to an AI system: the data, the system, and the business models	Traceability, explainability, open communication about limitations	• Purpose • Criteria of decision • Accuracy and metrics • Usage instructions • Limitations and risks
OECD Principles	2019	This principle is about transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes	Explainability, disclosure of relevant data	• Main factors in a decision, determinant factors, data, logic, or algorithm behind the specific outcome • How an AI system operates, is developed, trained, and deployed in the relevant application domain • Provide meaningful information and clarity about what information is provided and why
EU AI Act	2021		Communication, interpretability, explainability	• Identity and contact details of provider • Level of accuracy, robustness, and cybersecurity (Article 15) • Possible misuse, risks • Performance regarding person or groups the system is intended to be used • Risks and limitations • Input, train, validation, and testing data • Human oversight measures (Article 14) • Expected lifetime or necessary maintenance measures • Technical documentation (Annex IV) • Illustrative examples
JRC-Trustworthy Autonomous Vehicles	2021		Traceability, auditability, explainability, communication	• Communicating the users that they are interacting with an AI system • Inform about purpose, criteria, and limitations

descriptive elements, putting the main focus on ensuring that ML engineers are able to use these methods properly. A more compact approach to transparency reporting is the Transparency Index [25], a numerical indicator of how much relevant information, from an item list, the authors of the system have disclosed to the public. This is specifically designed for foundation models for Natural Language Processing (NLP) applications.

All these methods share the same objective: to provide useful, relevant, and critical information about an AI system or model, offering insights into its ethical and legal compliance, as well as its limitations and risks. In Table 2, we present a summary of the information each model-related reporting document includes.

Similar to models, methods for data reporting have also been proposed. Datasheets [26] accompany datasets and document their motivation, composition, collection process, and recommended uses, among other details. Additional data-related reports have emerged over time, the main ones being: Dataset Nutrition labels [27], Dataset Nutrition labels v2 [28], and Data Cards [29]. We propose an additional perspective to include for this communication requirement: MLOps Cards to describe MLOps activities. This type of report discloses information about the technical decisions and tools utilized to create the AI system covering its entire lifecycle (data gathering, labeling, training, testing, deployment, monitoring, etc.).

One limitation of the current approaches to model reporting is that all AI systems cannot be described in a standardized and general way. Each application field has its unique challenges, risks, and peculiarities. Depending on their intended use, AI systems may impact humans in various ways, and the nature of the task the model is designed to solve will determine its structure, making it challenging to consolidate all characteristics into a single standardized reporting method.

In this research, we will not delve into the creation of a Data Card specifically for CCAM. We believe that if a system aligns with the specifications of CCAM, the data used for its training will naturally align as well. Furthermore, it is important to note that data is an asset that may not necessarily be created by the same developers of a system. It might not even be created for a related purpose

TABLE 2 Sections of model-related reporting documents.

Model card [20]	Huggin face model card [22]	Factsheets [23]	Method cards [24]
• Model details	• Model details	• Overview	• Basic method information
• Intended use	• Uses	• Intended use	• Safety and troubleshooting
• Factors	• Limits and risks	• Data and model	• Data preparation
• Metrics	• Model training	• Performance	• Modelling and optimization
• Evaluation data	• Model evaluation	• Usage considerations	• Method benchmarking
• Training data	• Model examination		• Interpretability and explainability
• Quantitative analyses	• Environmental impact		• Robustness
• Ethical considerations	• Citation		
• Caveats and recommendation	• Technical specifications		
	• Model card contact		
	• How to get started		
	• Model card authors		
	• Glossary		
	• More information		

or the same field of application. This makes the need for a CCAM-specific Data Card more dispensable or less critical. Our Model Card has a section dedicated to data, where basic details can be provided. However, the presentation of a full Data Card is advisable for comprehensive understanding. Under our proposed methodology, any popular or generic Data Card can be used to complement the data-related information already included in the Model Card. While a CCAM-specific Data Card is not essential and a generic one is suggested, the presentation of a Data Card remains important.

2. Methodology

In this article, we advocate for the use of report documents as a method to enhance transparency in CCAM applications. The details of the AI system to be reported provide many ethical and technical definitions, contributing, at the same time, to other aspects of trustworthiness like accountability, governance, and robustness. This section describes how to create these documents, the criteria they adhere to, and the stakeholders involved in their creation and usage.

Our research aims to align and combine two main perspectives of AI applied to CCAM systems: the regulatory definition of AI and its requirements to be trustworthy. On the one hand, there has been significant progress in developing AI reporting documents that provide relevant information items. On the other hand, an adaptation to those requirements by the JRC provides a list of criteria for evaluating trustworthiness in CCAM. Both aspects face the necessity of increasing the transparency of AI systems. We aim to harmonize these two visions and suggest reporting documents as the answer to the communication requirements of the transparency normative with an additional focus on CCAM systems. The methodology presented in this article represents the first practical attempt, to the best of our knowledge, to merge these two converging work branches from two so far dissociated trends: AI regulation and CCAM, and AI reporting, as illustrated in Figure 1.

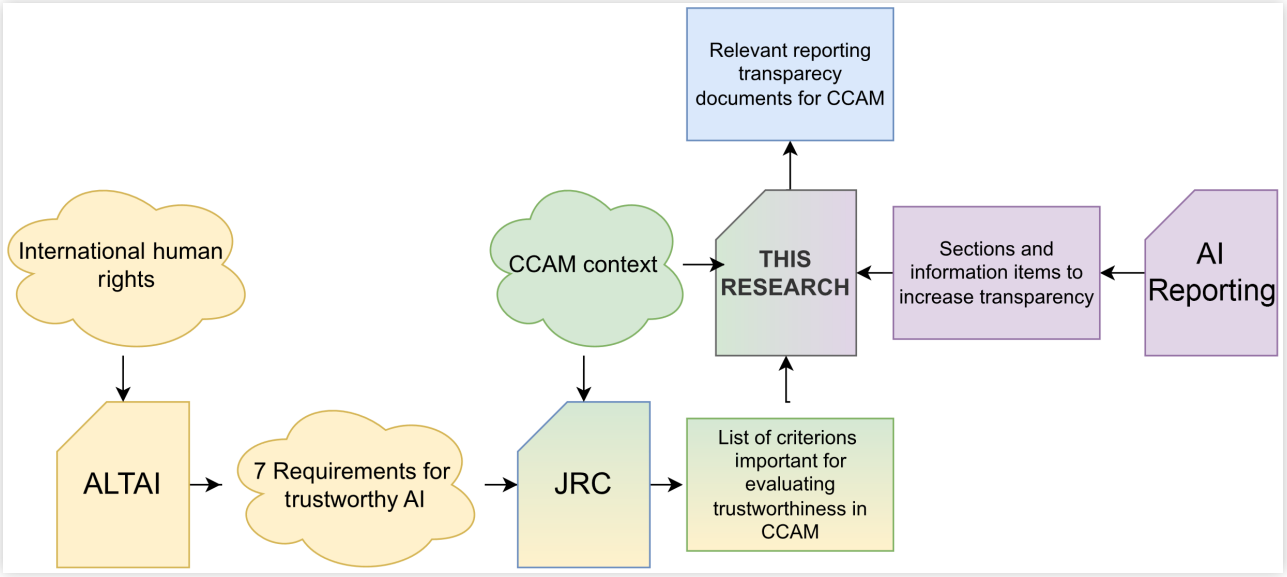
The proposed reporting cards in this article do not replace other advised methods of transparency, including technical documentation or logging systems (as required in the AI Act), that can be provided as complementary elements. Furthermore, these reporting cards should not be used as a reference or assurance for building trustworthy systems; they are merely reporting documents. The actual requirements and obligations are determined by current regulations, which may be subject to changes or updates.

Model Card

Looking forward to being compliant with the AI Act, we followed the criteria required for evaluating trustworthiness in CCAM published by the JRC. Our Model Card is inspired by the original concept from Mitchel et al. [20] and adapted to meet the requirements and specificities of CCAM applications. It keeps the exact name of Model Cards to align with the already established understanding of it as a reporting document of the system's information and to contribute to its standardization.

While the term "Model Card" may suggest a focus on individual ML models, our approach considers the system that uses, integrates, or contains the model as a whole. All documents of AI regulation or guidelines, including ours, refer to systems rather than isolated models, because the impact of a model cannot be fully understood without considering the system in which it is integrated. Analyzing an isolated model outside of its system context offers limited insights. The system interacts with users and performs specific functions; therefore,

FIGURE 1 Diagram of the methodology used for this research.



the model's impact is essentially the system's impact. For example, the impact of a driver distraction detection (DDD) system can vary significantly depending on whether it preprocesses or stores images of the driver. These processes could have privacy implications. It is important to note that these "preprocessing" and "storing" processes are part of the system, independent of the model. The system encapsulates the model along with these system-level processes to enable it to perform its function successfully. Therefore, when evaluating the impact of such a system, it is crucial to consider all its components in addition to the model itself.

The granularity and scope of the system to be reported are factors to be considered by the Model Card creator and specified within it. However, our Model Card design is intended to be ML model-oriented, mainly focusing on the part of the system that interacts directly with the model (i.e., input preprocessing, inference, post-processing, etc.). In case of conflicts with other models within the system, an additional Model Card should be provided. This approach ensures that each Model Card accurately reflects the role and impact of its corresponding model within the system.

Our Model Card template provides a structured format to guide developers in detailing comprehensive and accurate information about their models. This ensures that all pertinent information is presented in a clear and exhaustive manner, while also offering the flexibility to include additional details about the model as needed. The template, complete with sections and informational items for developers to fill out, is presented in Table 3.

To create the Model Card, we have compared its information items with the trustworthiness criteria (CR) provided by the JRC. This helped us make the appropriate changes in the Model Card to fit the needs of CCAM. Table 4 shows how each section in the Model Card matches with the requirement from the JRC (represented by its criterion ID), and their congruence with the seven KR of the ALTAI and the obligations described in the articles of the AI Act document. Additionally, it contemplates CCAM-related aspects like ODD's (Operation Domain Design), where the boundaries within which the automated driving system can safely operate or is designed to function are defined [31].

This analysis (correspondence between the items included in the Model Card and current ethical guidelines and regulations) is not meant to create a checklist for developers to follow and achieve legal compliance. Instead, it helps identifying key aspects of information related to ethical principles and regulations that are important to communicate to various users or stakeholders. Once a system meets these legal and ethical criteria, a descriptive summary of the system could include not only aspects that align with the communication and transparency recommendations stipulated in these documents (as shown in Table 1) but also other elements that demonstrate the system's compliance with the rest of the requirements. For instance, the "Ethical Considerations" section is intended for developers to express how they address these ethical requirements, but actual compliance with these requirements must be achieved beforehand.

MLOps Card

In addition to the Model Card, providing details about the development process of the model can offer further insights and enhance transparency. This highlights the

TABLE 3 Model Card template with sections and informational items.

1. Introduction		
1.1	**Author details** Information about the individuals or organizations responsible for developing the model.	**Information of the authors** • Name and the affiliations of the authors • Email of the authors or affiliation • Relevant links to websites • Affiliation logo • Participation of each of the authors in the creation
1.2	**Model info** A summary of the model, relevant information, and license.	**Brief description of the model** • Brief description of the model (its task, purpose, expected interactions with users, etc.) • Brief description of the model interaction with users • Layer of technology it belongs according to the JRC classification [16] • License of the model • Version and the creation date of the model • Links to repositories and/or website • Relevant scientific papers citations and links • Illustrative figure or an example of inference of the model
1.3	**Contact details** Contact information to report incidents or get more information about the model.	**Detailed contact information** • Name and the affiliations of the contact • Email address • Website or phone number
2. Model details		
2.1	**Model architecture** Detailed description of the model architecture, including hyperparameters, type of model, or other relevant details. Include diagrams or visualizations to help explain the architecture. ***A MLOps Card can be provided as a complement, which is highly advisable***.	**Description of the model architecture** • Type of model (e.g., neural network, decision tree) • Task of the model (e.g., classification, regression) • Number of layers or nodes • Information on the inputs and outputs of the model (e.g., images, multi-modal) • Information on the inputs and outputs of the system or the ones directly interacting with users • Diagrams or visualizations of the architecture • Description of the system the model is integrated into (e.g., Inference engine, data preprocessing or post-processing) • Describe the online continuous training process • Model updating process and criteria
2.2	**Training data** Detailed information about the data used to train the model. This can include source, size, preprocessing, data splits, and data augmentation. ***A Data Card can be provided as a complement, which is highly advisable.***	**Detailed information about the data used to train the model** • Purpose of the dataset (intended use) • Description of the dataset (e.g., size, format, number of instances, classes, features) • Source of the data (e.g., a public dataset, created data, simulated) • Statistics or visualizations to show the distribution of classes • Details of any preprocessing steps that were applied to the data **Information about the data splitting into training and validation sets** • Method used to split the data (e.g., random sampling, stratified sampling) • Additional information on the validation split if is different from the training split • Statistics or visualizations to show the distribution of classes in each set **Information on data augmentation techniques applied** • Data augmentation techniques used (e.g., rotation, flipping) • Examples of augmented data instances

(*Continued*)

TABLE 3 (Continued) Model Card template with sections and informational items.

2.3	**Evaluation data** Information about the data used to evaluate the model. This can include source, size, preprocessing. ***A Data Card can be provided as a complement, which is highly advisable.***	**Detailed information about the data used to evaluate the model** • Purpose of the dataset (intended use) • Description of the dataset (e.g., size, format, number of instances, classes, features) • Source of the data (e.g., a public dataset, created data, simulated) • Statistics or visualizations to show the distribution of classes • Details of any preprocessing steps that were applied to the data
3. Intended use and recommendations		
3.1	**Use cases and users** Description of the intended use for the model. This can include information about the problem the model is intended to solve, the target audience or users of the model, and any specific scenarios, situations, or users in which the model is intended to be used.	**Intended use cases of the model** • Description of the problem the model is intended to solve • Layer of technology it belongs according to the JRC classification (localization, scene understanding, path planning, control, user interaction) [16] • Level(s) of vehicle automation of the SAE [5] standard it is intended to be implemented • Technology readiness levels (TRL) of the system it is intended to be implemented • Risk classification according to the AI Act [15] • Examples of specific scenarios or situations in which the model is intended to be used (e.g., with illustrations, storytelling) • ODD description and specification of boundaries **Description of the developers to use the model (if any)** • Description of developer's limitations (location, private sector, public, government, etc.) • Level of expertise or technical knowledge required to implement the model • Other uses of the model (possible uses of the model besides the one described in this model card) **Description of the users that are analyzed or interact with the model (if any)** • Demographics (Information about the age, gender, location, etc. of the affected individuals or groups) • Relevant information for people with disabilities • Relevant information for children • Role of the users in the mobility context (e.g., driver, passenger, vulnerable road users) • Specified human-related assumptions and fixed values when building the model (human height, weight, Fitzpatrick skin types, age) • Expertise required to interact with the model
3.2	**Usage and recommendations** Detailed instructions and recommendations on how to use the model and what is not an expected behavior.	**Instructions on how to use the model (for developers)** • Brief description and details of where to find technical documentation • Instructions for developers to access the model for implementation • Guidance on how to use the model's output in downstream applications • Description of suggested ways to mitigate or work around any technical limitations • Guidance on how to use the model in an ethical and responsible manner **Instructions on how to use the model (for users)** • Explanation of how to interact with the model • Description of all intended human interactions with the model • Strategies implemented for human oversight • Description of how to exercise oversight • Methods of communicating the users that they are interacting with the system **Explanation on how to report unexpected behaviors** • Guidance on how to identify unexpected behaviors that need to be reported • Channel of communication to report incidences or incidents • Channel of communication to give feedback about user interactions

(*Continued*)

TABLE 3 (Continued) Model Card template with sections and informational items.

4. Evaluation performance		
4.1	**Metrics** Description of the specific metrics used to evaluate the model and any relevant thresholds or criteria for determining good performance and monitoring techniques.	**Performance metrics used to evaluate the model** • Accuracy to measure model performance • Precision and recall to measure performance in specific classes • F1 Score to measure performance on an unbalanced dataset • Metrics of overall performance (mission accomplished, per km failure rate, per km accident rate, # manual interventions) • Other AVs performance metrics like the ones suggested in JRC **Relevant thresholds or criteria for determining good performance** • Description of minimum acceptable values for each performance metric • Comparison of the model's performance to relevant benchmarks or baselines **Information about how the evaluation was performed** • Evaluation method used (e.g., cross-validation, holdout set) • Statistics or visualizations to show the model's performance • Real data/real scenarios evaluation procedures • ODD test coverage **Information about how the external evaluation was performed** • External parties' certifications or audits • Stakeholder participation in the evaluation • Description of user tests and user studies
4.2	**Results** Results of the performance evaluation, including any relevant statistics or visualizations are described in this section. This can include information about the overall performance of the model, as well as its performance on specific subsets of the data or scenarios.	**Results of the performance evaluation, including any relevant statistics or visualizations** • Explanation of each metric result • Report of overall performance using summary statistics (e.g., mean, median) • Visualization of performance in graphs or charts (e.g., confusion matrix, ROC curve) • Qualitative results and expert appreciations **Model's performance on specific subsets of data** • Performance in different demographic groups • Performance in different situations or scenarios • Performance on simulated data, test tracks, and real traffic situations • Report of performance for defined ODD
4.3	**Monitoring** Description of strategies and metrics to monitor of the performance of the model after deployment.	**Strategy of continuous monitoring of the performance** • Explanation of the strategy of continuous monitoring • Measurement of concept drift • Measurement of data drift • Description of periodical tests and maintenance
5. Limitations and risks		
5.1	**Limitations** Information about any limitations or restrictions on the use of the model. This can include technical limitations as well as ethical or legal limitations.	**Limitations or restrictions on the intended use of the model** • Technical limitations (e.g., the model only works with certain types of data) • ODD-related limitations and boundaries (e.g., the model only works with certain scenarios) • Reporting of any known failures of the system • Specification if it is sensible to occlusions • Specification if a calibration step is needed • Ethical or legal limitations (e.g., the model should not be used for certain purposes, locations) • Restrictions on the type or format of input data **System response when outside the ODD** • Description of the fallback plan • Description of system behavior when failure or outside the ODD • Description of system behavior when close to the limits of ODD • Description of strategies for handling low confidence scores or model uncertainty

(*Continued*)

TABLE 3 (Continued) Model Card template with sections and informational items.

5.2	**Risks and impact** Risks that can be related to the usage of the model.	**Identification of the risks** • Description of the risk assessment • Risk of attachment, addiction, and user behavior manipulation • Risks from robustness deficiencies • Risks from lack of explainability • Risks from bias and discrimination • Risks from vulnerability to attacks • Privacy concerns **Impact of the model (positive or negative)** • Impact on end users' life • Impact on society • Environmental impact **Explanation of the adherence to the ALTAI** • Description of compliance with the key requirements
6. Ethical considerations		
6.1	**Fairness** Discussion of any potential biases in the model or its intended use. This can include information about how the model was tested for fairness, as well as any steps taken to mitigate potential biases.	**Model's fairness using appropriate fairness metric** • Statistical parity difference to measure group fairness • Equal opportunity to measure individual fairness • Other fairness metrics **How any potential biases are mitigated** • Reweighting of the training data to balance underrepresented groups • Modified loss function to weigh the loss contribution of the classes • Other used bias mitigation techniques **Other remaining possible biases** • Description of how representative the train and test data used for the intended use case • Diversity of end users included in the data • Diversity of users testing and validating the model • Description of inclusion in the analysis of people with disabilities • Description of inclusion in the analysis of children
6.2	**Privacy** Information about how the model handles user data and any privacy considerations. This can include information about what data is collected, how it is stored and used, and any measures taken to protect user privacy.	**What data is collected, how it is stored and used** • Data map showing what data is collected and how it flows through the system • Description of the data retention policy and how long data is stored • Explanation of how personal data is used and shared with third parties **Implemented measures to protect user privacy** • Data anonymization techniques to de-identify user data (blurring, pseudo anonymization) • Person blurring and license plate blurring • Encryption of user data at rest and in transit • Implementation of access controls to data • Federated learning techniques **Relevant privacy regulations and standards compliance** • Link to consent and privacy agreements • Description of how the user is asked for consent • Protocol to the user to ask for deletion of their data • Communication channel to the user to know their personal data treatment • Explanation of compliance with the GDPR [30]
6.3	**Explainability** Information about how decisions made by the model can be explained and reviewed. This can include information about the explainability of the model.	**Used explainable AI (XAI) methods to increase model explainability** • Description of explainable AI techniques to make the model's decisions more explainable (e.g., saliency maps) • Uncertainty measurement • Description of the logging system • Other used XAI techniques

TABLE 4 Correlation between the model card and ethical guidelines and requirements.

Model card item	JRC criterion (CR)	ALTAI requirement	AI act article
1.1 Authors details	7.1, 7.7	KR7	11(Annex IV), 13
1.2 Model info	4.5	KR5	11(Annex IV), 13
1.3 Contact info	7.1, 7.7	KR7	11(Annex IV), 13
2.1 Model architecture	1.7, 2.7, 2.22, 4.5	KR1, KR2, KR4	11(Annex IV), 13, 14
2.2 Train data	2.14, 2.16	KR2	10, 11(Annex IV)
2.3 Eval data	2.14, 2.16	KR2	10, 11(Annex IV)
3.1 Use case and user	2.2, 2.16, 6.7	KR2, KR6	6, 11(Annex IV), 13
3.2 Use and recommendations	1.2, 1.8, 2.2, 4.3, 4.4, 4.5	KR1, KR2, KR4	11(Annex IV), 13, 14
4.1 Metrics	2.3, 2.17, 4.5, 5.10, 7.2, 7.7, 7.3	KR2, KR4, KR5, KR7	9, 11(Annex IV), 15
4.2 Results	2.13, 2.15, 2.19, 5.10	KR2, KR5	9, 11(Annex IV), 13, 15
4.3 Monitoring	2.10, 2.15	KR2	11(Annex IV), 15
5.1 Limitations	2.9, 2.11, 2.18, 2.20, 2.21, 4.5	KR2, KR4	11(Annex IV), 13
5.2 Risks and impact	1.1, 1.6, 2.8, 2.12, 2.13, 3.2, 5.9, 6.1, 6.8, 7.6	KR1, KR2, KR3, KR5, KR6, KR7	9, 11(Annex IV), 13, 14
6.1 Fairness	5.1, 5.2, 5.4, 5.5, 5.6, 5.7, 5.8	KR5	10, 11(Annex IV)
6.2 Privacy	2.3, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6	KR2, KR3	10, 15
6.3 Explainability	4.1, 4.2	KR4	12

importance of reporting the software infrastructure and processes involved in the model's creation, implementation, and monitoring. Hence, we propose an MLOps Card in this research, a document that outlines the technological decisions made during the development of an AI system, acting in the same time as technical documentation. This document can help describe how the monitoring of the model is being handled or how the CI/CD (continuous integration and continuous delivery/deployment) good practices are implemented.

This card features a diagram that represents the processes and tools used in an ML solution, considering the different stages of its lifecycle. The diagram in Figure 2 primarily consists of four main stages (commonly involved when creating an AI model): data collection, training, testing, and deployment. Inside each one, there are common subtasks or processes performed when building an ML model, represented by blocks. The ones presented in the diagram are optional, and others can be included.

In addition, there are other processes that support the construction of the model throughout its entire lifecycle. These are represented at the bottom of the diagram; these are aspects such as computing, virtualization, storage, and visualization. An Orchestrator is positioned at the top left of the diagram to indicate that an orchestrion method of the stages or some process is implemented to automatize its execution. The MLOps Card is designed to enable the developer to specify the tool or method they used for each block at each stage of the process. The diagram also includes arrows indicating the flow of the lifecycle and the connection between blocks. These arrows can be edited to represent an iterative process or a one-pass development. This allows for flexibility in representing the flow of the model's lifecycle and the connections between the different stages.

The proposed MLOps Card provides a unified way of representing the technical operations of an AI solution. This way, the card allows for easy identification of the flow of the lifecycle of the model and the compatibility between the used tools. It also provides information about the sources of the data, the annotation strategies, the storage of the model, among other technical choices. An MLOps Card can be provided to complement the 2.1, Model Architecture Section of the Model Card.

Stakeholders

Due to the inherent complexity of CCAM systems and services, multiple stakeholder roles participate in the creation, deployment, or utilization of AI-based systems. The developer is often seen as the main role, but other roles are responsible for different subprocesses such as testing, management, and quality assessment. Analogously, the final user who is interested in these documents can be a driver of a bus that uses a CCAM system, but it could also be a passenger of that bus or the manager of the company that owns the bus. Each role will have different interests and the AI-based system will affect them or interact with them in a different way.

For the sake of simplicity, we define three main stakeholder profiles: user, auditor, and developer. Figure 3 illustrates these profiles and how the reporting documentation can interest each of them. (i) **User:** A person who interacts with the AI system or is affected by it somehow. The user could be or not be a connoisseur of technology so that information can be provided about the non-technical aspects of the system, like the purpose, limitations, and risks. (ii) **Auditor:** This refers to any role in charge of verifying or validating the system. The Auditor's perspective might be more technical, oriented to testing metrics,

FIGURE 2 MLOps diagram.

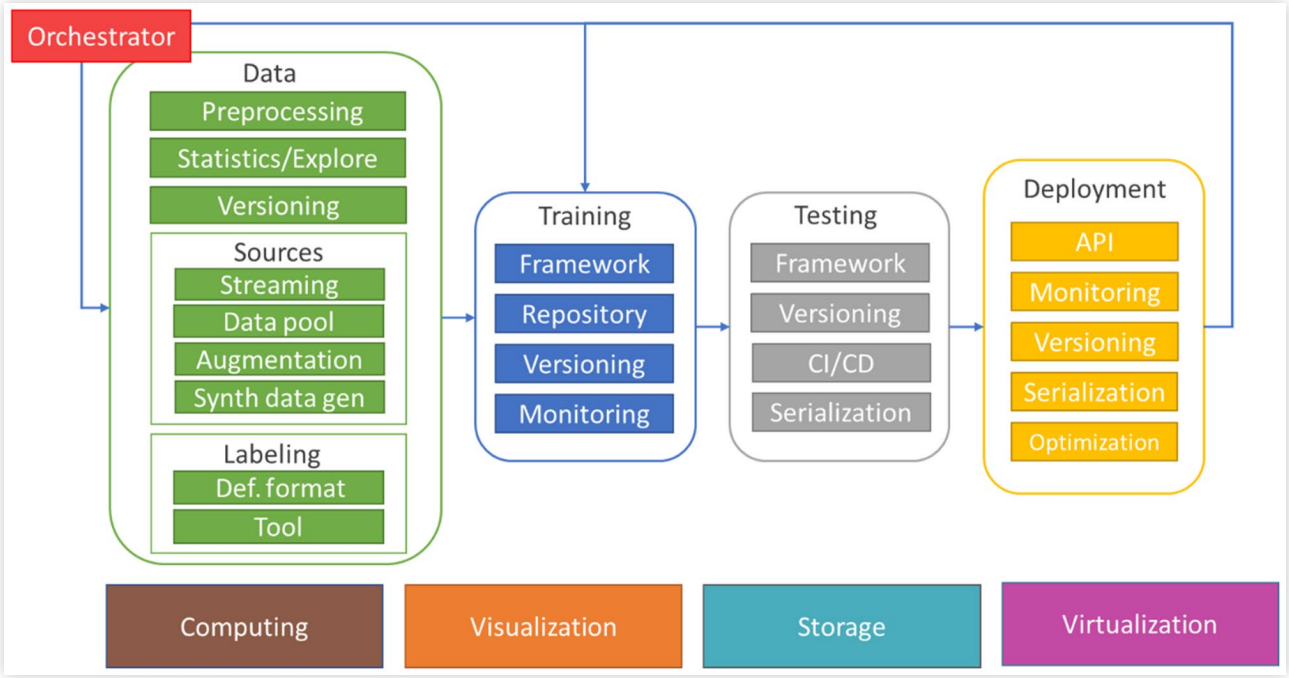
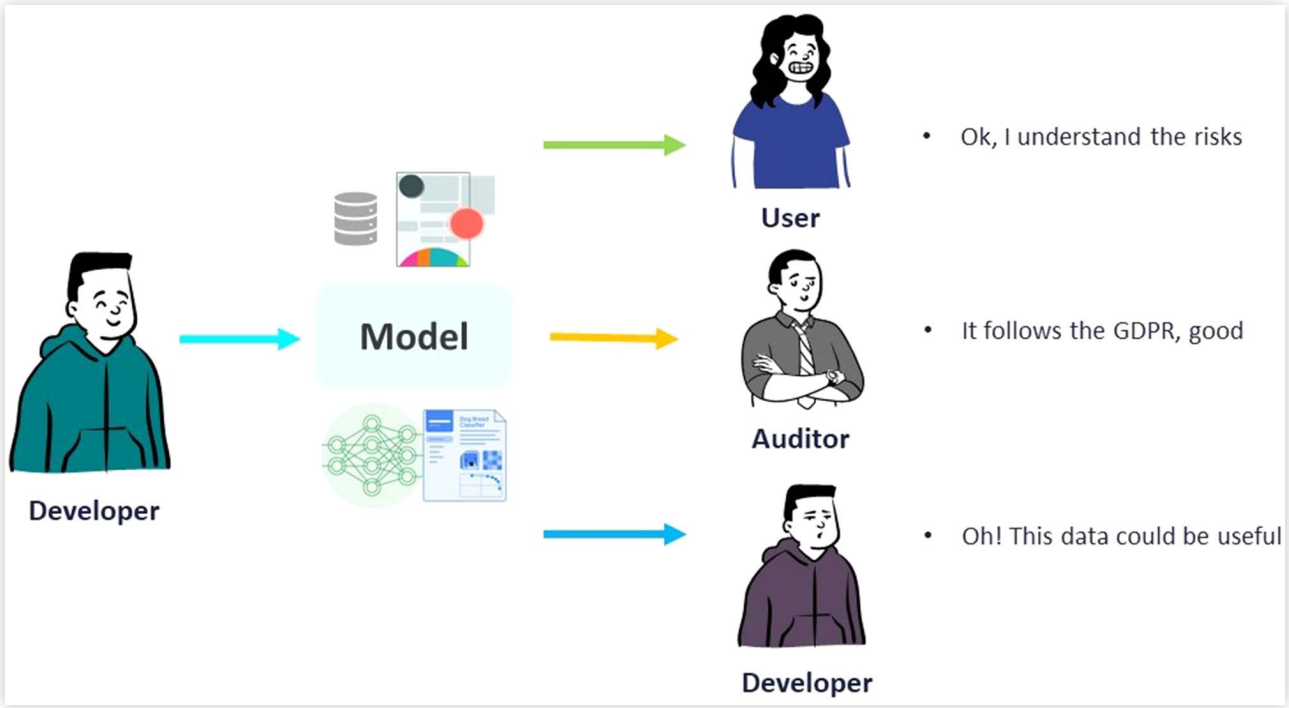


FIGURE 3 Representation of the stakeholders of the reporting documents.



ethical metrics, certifications, compliance with regulatory requirements, and the completeness of the reporting document itself. (iii) **Developer:** This refers to roles participating in any step of the process of creating an AI system. This, too, is a stakeholder of a reporting document to discover good practices, compliant data, and models. On the other hand, the developers are responsible for building the report documents; independent of how long the list of developers is, there must be a unified and unique report document accompanying the AI system. The person(s) who acknowledges the system's authorship should also be the representative to ensure it meets the transparency requirements.

The AI Act identifies some additional stakeholders in Article 3. These include providers, small-scale providers, users, authorized representatives, importers, distributors, and operators. All of these, except for users, can match our definition of the developer. Providers have the obligation to comply with the requirements, and as specified in Article 16, to prepare the technical documentation and keep it updated. While the users shall use the system in accordance with the instructions of use.

In the context of CCAM, identifying the user is a challenging task. The JRC document outlines multi-user considerations, indicating that numerous actors must be treated independently, such as the driver of the ego-car, other drivers, passengers, or vulnerable road users (VRUs). The target users shall be identified as the ones directly affected by the system, in order to facilitate effective communication with them. A system operating in a traffic situation impacts all road actors: VRUs, drivers, and passengers. On the other hand, other developers also serve as users of the reporting document. This is easier to address since they are expected to have a technical background or be interested in information within a context where they have specific expertise.

3. Discussion

CCAM is not the first industry to face transparency and communication requirements. The food and pharmaceutical industries have faced the same. Nutritional facts tables are attached to any commercial product, as well as a drug specification leaflet in every pharmacy product. One key valuable lesson to be learned from these industries when communicating with users is the importance and urgency of standardization of communication strategies. It ensures a common understanding and makes it easier for users to find and comprehend the necessary information.

One main concern of disclosing information is the potential risks and problems related to intellectual property and the security of the system. These reasons are the main drawback to developers when preparing the proposed set of reporting documents. More guarantees must be provided to developers and their intellectual properties, while the information that could generate a security problem should be reserved. More discussions and research should be performed regarding this matter.

The format of reporting information is another area that warrants research. These cards should not only be available and accessible to all, but mechanisms should also be in place to be able to keep them updated. How could these cards be presented in a multi-user scenario? Can a passenger of a bus with an integrated AI system access this information? And would it be in a different way than the professional driver of that bus?

Similarly, developing more abstract levels of communication, such as indexes, labels, color coding, badges, scales, and the like, could provide a more concise and quick way to share the level of trustworthiness or other relevant information with the user, achieving a more informed and confident environment of coexistence with AI systems.

4. Conclusions

The proposed set of reporting documents, shaped as cards, presents a novel approach to fulfilling AI transparency requirements, contributing to the first steps to fostering trust in AI systems in CCAM applications. Our proposal acknowledges that each AI system deserves a tailored approach on how to apply the ethical guidelines due to the sector's unique characteristics and challenges. We adapt the existing Model Card to the CCAM needs, showing the correlation of the information items with different criteria, requirements, and obligations for trustworthiness.

While these cards may contain extensive and potentially sensitive information that developers might hesitate to disclose, a standardized process for communicating the risks that addresses concerns associated with an AI system needs to be defined, established, and followed by all actors participating in the development and production. In this article, we introduce a methodology that is ready for all stakeholders in the CCAM industry, governmental entities, and potential users to explore, implement, and utilize. By doing so, these reporting cards can be evaluated under varying degrees of technical specificity that developers may wish to communicate, or for describing different types of systems. Their interests and opinions shall be gathered for an effective and specialized implementation, concerning the scope and format of the reporting documents, aiming to present them to the end users in a collaborative environment.

We believe that the introduction of AI systems into society is beneficial, as they are powerful tools that can greatly assist humans. It is within our capability to understand them and use them effectively. We should remain

committed to transparency, ethical practices, and the pursuit of trustworthiness in AI. For this, Data, Model, and MLOps Cards can be used as a starting point within the CCAM domain.

Acknowledgements

This work was funded by the Horizon Europe programme of the European Union, under grant agreement 101076754 (project AITHENA). Views and opinions expressed here 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.

Contact Information

Paola N. Cañas
pncanas@vicomtech.org

Marcos Nieto
mnieto@vicomtech.org

References

1. Thadani, T., Rachel, L., Imogen, P., Faiz, S. et al., "Tesla Autopilot Crash Analysis," *The Washington Post*, October 6, 2023, accessed December 22, 2023, https://www.washingtonpost.com/technology/interactive/2023/tesla-autopilot-crash-analysis.
2. Shepardson, D., "GM's Cruise Recalling 950 Driverless Cars after Pedestrian Dragged in Crash," Reuters, November 8, 2023, accessed December 22, 2023, https://www.reuters.com/business/autos-transportation/gms-cruise-recall-950-driverless-cars-after-accident-involving-pedestrian-2023-11-08.
3. Drapkin, A., "ChatGPT and AI Scams to Watch Out for and Avoid," Tech.co, July 25, 2023, accessed December 22, 2023, https://tech.co/news/chatgpt-ai-scams-watch-out-avoid.
4. Pilici, S., "Beware of the Viral 'MrBeast 10,000 iPhone 15 Pro for $2' Scam," Malwaretips, October 30, 2023, accessed December 22, 2023, https://malwaretips.com/blogs/mrbeast-10000-iphone-15-pro-for-2-scam.
5. SAE International, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," J3016 202104, 2021.
6. Floridi, L., Cowls, J., Beltrametti, M. et al., "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds and Machines* 28, no. 4 (2018): 689-707.
7. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*, 1st ed. (Piscataway, NJ: IEEE, 2019)
8. Winfield, A.F.T., Serena, B., Louise, A.D., Takashi, E. et al., "IEEE P7001: A Proposed Standard on Transparency," *Frontiers in Robotics and AI* 8 (2021): 665729, doi:https://doi.org/10.3389/frobt.2021.665729.
9. Hagendorff, T., "The Ethics of AI Ethics: An Evaluation of Guidelines," *Minds and Machines* 30, no. 1 (2020): 99-120, doi:https://doi.org/10.1007/s11023-020-09517-8.
10. Rességuier, A. and Rodrigues, R., "AI Ethics Should Not Remain Toothless! A Call to Bring Back the Teeth of Ethics," *Big Data & Society* 7, no. 2 (2020): 2053951720942541.
11. Mittelstadt, B., "Principles Alone Cannot Guarantee Ethical AI," *Nature Machine Intelligence* 1, no. 11 (2019): 501-507, doi:https://doi.org/10.1038/s42256-019-0114-4.
12. High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy Artificial Intelligence," European Commission, Brussels, 2019.
13. High-Level Expert Group on Artificial Intelligence, "Assessment List for Trustworthy Artificial Intelligence," European Commission, Brussels, 2020.
14. OECD, "Recommendation of the Council on Artificial Intelligence," OECD/LEGAL/0449, 2019.
15. European Commission, "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on AI (Artificial Intelligence Act) and Amending Certain Union Legislative Acts," COM/2021/206 Final, Brussels, April 4, 2021.
16. Fernández Llorca, D. and Gómez, E., "Trustworthy Autonomous Vehicles," EUR 30942 EN, Publications Office of the European Union, Luxembourg, 2021, doi:https://doi.org/10.2760/120385.
17. Borg, M., Joshua, B., Linus, C., Fredrik, O. et al., "Exploring the Assessment List for Trustworthy AI in the Context of Advanced Driver-Assistance Systems," in *2021 IEEE/ACM 2nd International Workshop on Ethics in Software Engineering Research and Practice (SEthics)*, Madrid, Spain, 2021, 5-12, https://doi.org/10.1109/SEthics52569.2021.00009.
18. Jobin, A., Ienca, M., and Vayena, E., "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence* 1, no. 9 (2019): 389-399, doi:https://doi.org/10.1038/s42256-019-0088-2.
19. Li, B., Qi, P., Liu, B., Di, S. et al., "Trustworthy AI: From Principles to Practices," *ACM Computing Surveys* 55, no. 9 (2023): 1-46, doi:https://doi.org/10.1145/3555803.
20. Mitchell, M. et al., "Model Cards for Model Reporting," in *FAT'19: Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, January 29–31, 2019, Association for Computing Machinery, New York, 10 pages, https://doi.org/10.1145/3287560.3287596.

21. Hugging Face, "Model Card Examples," n.d., accessed December 22, 2023, https://huggingface.co/docs/hub/model-card-appendix#model-card-examples.
22. Hugging Face, "Appendix A: User Study," n.d., accessed December 22, 2023, https://huggingface.co/docs/hub/model-card-appendix#appendix-a-user-study.
23. Arnold, M., Bellamy, R.K.E., Hind, M., Houde, S. et al., "FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity," *IBM Journal of Research and Development* 63, no. 4/5 (2019): 6:1-6:13.
24. Adkins, D., Alsallakh, B., Cheema, A., Kokhlikyan, N. et al., "Method Cards for Prescriptive Machine-Learning Transparency," in *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI (CAIN)*, Pittsburgh, PA, 2022, 90-100, Association for Computing Machinery, New York, https://doi.org/10.1145/3522664.3528600.
25. Bommasani, R., Klyman, K., Longpre, S., Kapoor, S. et al., "The Foundation Model Transparency Index," arXiv [Cs.LG], 2023, http://arxiv.org/abs/2310.12941.
26. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W. et al., "Datasheets for Datasets," *Communications of the ACM* 64, no. 12 (2021): 86-92.
27. Holland, S., Hosny, A., Newman, S., Joseph, J. et al., "The Dataset Nutrition Label: A Framework to Drive Higher Quality Data Standards," arXiv, May 2018.
28. Chmielinski, K.S., Newman, S., Taylor, M., Joseph, J. et al., "The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence," arXiv, January 2022.
29. Pushkarna, M., Zaldivar, A., and Kjartansson, O., "Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI," arXiv [Cs.HC], 2022, http://arxiv.org/abs/2204.01075.
30. European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)," OJ L 119, 2016, 1-88.
31. International Organization for Standardization, "Road Vehicles—Test Scenarios for Automated Driving Systems—Specification for Operational Design Domain," ISO 34503:2023, 2023, https://www.iso.org/standard/78952.html.