Towards Trustworthy AI in the Public Transport Domain*

Giuseppe Riccardo Leone[1,*], Andrea Carboni[1], Giulio Del Corso[1], Silvia Gravili[1], Davide Moroni[1], Maria Antonietta Pascali[1] and Sara Colantonio[1]

[1]*Institute of Information Science and Technologies "Alessandro Faedo", National Research Council of Italy (ISTI-CNR)*

Abstract

In the context of rapidly evolving urban landscapes, the demand for enhanced mobility services has become increasingly critical. Traditional transportation systems struggle to keep pace with the growing complexity of commuting patterns and the diverse needs of urban residents. While AI can play a strong role in addressing these emerging demands, a parallel need for trustworthy services is also arising, which must be adequately met to ultimately provide equitable and ethical services to society. Based on these considerations, we explore the relevant dimensions of AI trustworthiness and propose how they can be transferred and demonstrated in a large-scale pilot focused on public transportation and exploiting advanced visual analytics paradigms based on pervasive computing. To this end, we present the FAITH risk management framework, ongoing activities, and preliminary results towards its implementation in the pilot project.

Keywords

Trustworthy AI, Computer vision, Intelligent Transport System, Edge computing, Privacy by design

1. Introduction

In today's rapidly evolving urban landscapes, the demand for improved mobility services has become a critical need. As cities expand and populations grow, traditional transportation systems struggle to keep pace with the increasing complexity of commuting patterns and the diverse needs of residents. Traffic congestion, environmental concerns, and the inequitable distribution of transportation resources highlight the urgency for innovative solutions. Improved mobility services are not just about enhancing efficiency; they play a vital role in fostering social equity, economic growth, and environmental sustainability. By embracing a holistic approach to mobility –incorporating public transit, ride-sharing, walking, biking, and emerging technologies– societies can create interconnected systems that promote accessibility, reduce carbon footprint, and improve the quality of life for all citizens.

Artificial Intelligence (AI) holds immense potential, in general, to revolutionize mobility services. There are several key roles AI can play in the public transport domain: **Optimization of Transportation Networks** - by analyzing real-time data from traffic patterns and user demand, AI can optimize routes and schedules, reduce wait times, minimize congestion, and enhance the efficiency of public transports. **Personalized Mobility Solutions** - by offering personalized transportation options based on user preferences, needs, and behaviors. **Predictive Analytics** - by leveraging historical data and patterns, AI can predict future transportation trends and demands. **Enhance Safety and Security** - by analyzing patterns of accidents or identifying unsafe conditions. **Environmental Sustainability** - by optimizing energy use in transportation systems prioritizing eco-friendly routes boosting the use of electric vehicles.

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

*This work was supported by 'Fostering Artificial Intelligence Trust for Humans towards the optimization of trustworthiness through large-scale pilots in critical domains' (FAITH) project, funded by the European Union's Horizon Programme under grant agreement No 101135932.

*Corresponding author.

giuseppericcardo.leone@cnr.it (G. R. Leone)

0000-0002-7441-8528 (G. R. Leone); 0000-0003-3227-3487 (A. Carboni); 0000-0003-4604-2006 (G. D. Corso); 0000-0002-8761-2709 (S. Gravili); 0000-0002-5175-5126 (D. Moroni); 0000-0001-7742-8126 (M. A. Pascali); 0000-0003-2022-0804 (S. Colantonio)

In summary, AI has the potential to create a more connected, efficient and equitable mobility landscape: technological advances will enable all citizens have access to reliable transportation options, fostering inclusive urban environments. However, its widespread impact across various societal domains – including mobility– has led to growing public awareness and concern regarding the potential misuse of AI, as well as a demand for ethical and trustworthy systems. There is a pressing need for scientifically grounded and holistic strategies to enhance human trust in AI. This is the core objective of the FAITH project (Fostering Artificial Intelligence Trust for Humans)[1], which will be accomplished through the development of the FAITH Artificial Intelligence Trustworthiness Assessment Framework (FAITH AI_TAF): this methodology will be instantiated, demonstrated, and refined via a representative selection of large-scale pilots in critical domains, including mobility.

In this paper, we aim to introduce the framework in the context of the transportation pilot, providing the following main contributions: 1) exploring the aspects of trustworthiness in the transit domain, centered around the use of deep learning models for computer vision, encompassing object analysis as well as human activity recognition; 2)investigating the use of pervasive artificial intelligence methods for pilot deployment, analyzing both their design prerequisites and the potential threats that may emerge throughout the entire life-cycle; 3)reporting preliminary findings and initial feedback to further promote trust in AI within the specific pilot application domain. The remainder of this paper is organized as follows: Section 2 surveys related works covering both the regulatory framework and the concept of trustworthiness; Section 3 is devoted to the description of the FAITH approach to risk assessment with major details on the transport domain; Section 4 presents insights from the activities of the first project's year, while Section 5 concludes the paper and suggests potential directions for future research.

2. Related Work

From the previous brief survey, it is evident that Artificial Intelligence is expected to play a pivotal role in the ongoing digital transformation of public transport. Most academic and industrial applications have focused on performance and usability, measuring success through recognition rates or task improvements achieved via AI, while key features such as safe, conscious use, and societal acceptance are more difficult to define, measure, and verify. The development and deployment of AI systems should align with ethical guidelines, which means considering the broader societal impacts and ensuring that AI benefits humanity as a whole[2]. In response to these needs, various global, international, and national initiatives have been established to provide guidelines and requirements for AI trustworthiness. Among them are deliverables from the High-Level Expert Group on Artificial Intelligence [3], the proposed European Union Artificial Intelligence Act (EU AI Act) [4], the White House Blueprint for an AI Bill of Rights [5], and the NIST AI Risk Management Framework (RMF) [6] for risk assessment. Additionally, several efforts aim to regulate and establish best practices for the complex interplay between cybersecurity and AI, such as those promoted by the EU Agency for Cybersecurity (ENISA) [7].

The ISO/IEC standard TS 5723:2022 [8] defines trustworthiness as the "ability to meet stakeholders' expectations in a verifiable way". Similarly, NIST views trustworthiness as an objective attribute of a system, stating that "trustworthiness of a system is based on the concept of assurance" [9]. In the context of communication and systems [10, 11], trustworthiness is a multifaceted concept that, when applied to the AI domain, covers several key areas: **Transparency and Explainability**: Transparency and explainability entail fully documenting the entire lifecycle of an AI system, to allow the user know how decisions are made, including the data and algorithms used. They also helps build trust by allowing stakeholders to see the rationale behind AI outputs. **Fairness**: AI should be designed to minimize bias and ensure equitable treatment across different groups [12]. This requires careful consideration of training data and algorithms to avoid reinforcing existing prejudices or creating unfair outcomes. **Accountability**: There must be clear lines of responsibility for AI decisions. Stakeholders should know who is accountable for the actions of an AI system, especially in cases of harm or unintended consequences. **Reliability**: Trustworthy AI systems must perform consistently under

varying conditions. This includes robustness to adversarial attacks and resilience to unexpected inputs, ensuring that the AI operates correctly across a range of scenarios. **Privacy**: Protecting user data is crucial for trust. As any system using sensitive data (e.g., visual data), the AI systems should incorporate strong data protection measures, ensuring that personal information is handled ethically and securely. **Stakeholders Engagement**: Involving users in the design and evaluation of AI systems can enhance trust. This means soliciting feedback, addressing concerns, and involving diverse perspectives in the development process. **Compliance and Governance**: Adhering to legal regulations and industry standards is essential for maintaining trustworthiness. Organizations should establish governance frameworks to oversee AI deployment and ensure compliance with relevant laws.

Trustworthiness must therefore be considered across multiple dimensions of analysis and systematically established and addressed throughout the entire life-cycle of an AI system. To take a holistic view of trustworthiness at all stages, it's crucial to recognize the distinct risks that AI poses compared to traditional ICT systems. In addition to addressing each aspect of AI trustworthiness, it is essential to consider their interactions and trade-offs, as these remain underexplored yet crucial for real-world applications. For example, the need for data privacy may create tension with the desire to provide detailed explanations of system outputs, and pursuing increased accuracy of an AI/ML system may reduce its explainability [13]; however, approaches such as "robustness to explainability" (to improve robustness thanks to explainability) and "Robustness to fairness" (to avoid bias and discrimination in all aspects of the AI ecosystem) have been reported [14]. In any case, improving individual elements of trustworthiness in isolation does not guarantee a more trustworthy or effective system, but carefully balancing competing goals and synergies should be sought.

3. Methodology

The approach devised in FAITH is to build a general framework for the assessment of trustworthiness, the "FAITH AI_TAF", feasible enough to be versioned in several specific application domains, possibly critical, in order to enable testing, measurement, and optimization of risks associated with AI trustworthiness. Such a framework builds upon NIST AI RMF, addressing EU legislative requirements and ENISA guidelines to ensure trustworthiness by design.

3.1. The Large-Scale Pilots

In more details, FAITH has identified seven challenging application domains (i.e., media, transport, education, robotics, industry, healthcare, and wellbeing) in order to put in action such a theoretical framework, and test it in Large-Scale Pilots (LSPs). The general scheme is made of two main steps, to be run in an agile manner, and they are: 1) the definition of the main characteristics of trustworthiness for an AI-based system, i.e., "dimensions"; 2)the identification of tools (available, or to be designed and developed) and suitable metrics (existing or novel) useful to assess each dimension.

The agile approach means that each step has to be performed leveraging the contribution of all the involved parties: AI modelers and AI users, stakeholders and domain experts, and even citizens; of course, such a contribution can have a dominant character of co-creation, investigation, validation, verification or fine-tuning, depending on the development stage the framework is in.

In action, this means that, after agreeing on a tentative list of trustworthiness dimensions and the tools for assessing them, in each pilot, an AI-based system should be put under the magnifying glass to check if such dimensions are all relevant to assess trustworthiness or, on the other hand, if new ones have to be added to the general framework; and if suitable assessment procedures and tools are already available and used or not. The evaluation of trustworthiness may affect any stage of the AI life-cycle. Each pilot is composed of three phases: (i) initial, (ii) replication, and (iii) post-project. Phase 1 will setup all the main procedures and solutions for an instance of the domain, covering the entire AI life-cycle. Phase 2 will build on the results of Phase 1, leveraging insights and lessons learned to refine and expand upon the initial solutions; this staged approach demonstrates the reusability and adaptability of the project's solutions, addressing similar goals as the initial phase but with added complexity and

real-world applicability. The post-project phase (Phase 3) aims at ensuring the long-term sustainability of project outcomes: impact and uptake will be maximized through external end-users engagement, technology transfer, incubation, and innovation management.

3.2. The LSP on public transportation

This LSP aims to deliver a scalable, privacy-preserving platform utilizing pervasive AI and video analytics with the goal of enhancing safety, reliability, efficiency, and cleanliness, on board public transportation thereby positively impacting citizens' lives. The system will analyze closed-circuit video streams collected on board trains during daily operations, focusing on identifying garbage, unattended objects, equipment, furniture, missing or damaged items, available seats, passenger counts, and safety issues. Privacy-by-design principles will be implemented through visual anonymization, the data will be processed locally with edge computing, and avoiding biometric data collection to alleviate concerns about invasive surveillance. Only synthesized information will be sent to an operations center, providing managers with actionable insights in an "augmented intelligence" mode.

The research team of ISTI-CNR, which is the leader of this LSP, has strong expertise in the Intelligent Transportation System (ITS) domain. In particular, the application of computer vision to public transport has been faced in the SPACE project [15]: a network of cameras and embedded systems was employed to build a scalable, edge-computing solution for the pervasive monitoring onboard carriages and in stations. Computer vision models, leveraging state-of-the-art deep learning paradigms for real-time object detection and tracking, were integrated to assess human activity and detect potential threats, such as fires, unattended luggage, and acts of vandalism. Additionally, a human activity recognition module was included to detect fights or disturbances [16]; This previous work is the starting point for the activities to be carried out in the pilot; therefore, the system will incorporate integrated video analytics services based on deep learning for image analysis, object detection, scene analysis, and activity recognition. These services will be deployed on-board edge computational nodes, preventing video transfer to remote locations to reduce security risks. The data collection will be extensive and pre-screened using existing algorithms, allowing for refinement and the potential addition of new algorithms for behavior characterization (e.g., loitering) and safety/security concerns (e.g., fall detection). The output from these algorithms will enable cross-correlation of mobility patterns based on data from single cameras.

Successful real-world implementation depends on stakeholders trusting the system to be technically robust, accurate, and reliable. This includes protection against malicious attacks and system failures that could jeopardize public safety.

As explained in the previous section, the pilot will progress through three iterations. The initial phase is for deploying distributed AI-based video analytics at a regional level, using around 100 edge nodes with some virtualized nodes, starting with reserved coaches of the Italian railway operator Trenitalia in selected regions such as Tuscany, Lazio, or Apulia. The replication phase will expand usage across three different areas, including train coaches and stations, with a daily passenger flow exceeding 10,000. During the post-project phase, the scale is intended to be distributed all over the country, deploying at least 1,000 edge nodes, either physical or virtualized.

4. Preliminary activities

The FAITH project started in January 2024. While it is still too early to have validated experimental results, significant activities have been carried out on the following topics:

Risk Profile - Discussion on the dimensions of interest and measurement methods are a focal point to define the risk profile of the domain. Privacy and Data Protection is one of the most important. The process of legal evaluation of the impact of the proposed software solutions has begun, suitability will be assessed by the CNR Ethical Committee to which we will submit the Data Protection Impact Assessment (DPIA), which is required under the GDPR for any new project that is likely to involve "a high risk" to other people's personal information.

Stakeholders - The engagement of stakeholders will include: a) public transport operator managers: they will provide access to facilities and vehicles, guiding technical and non-technical requirements; b) public transport planners: they will gain knowledge for optimizing lines, frequency, and service; c) data scientists: engaged throughout the AI management life-cycle; d) passengers: as end users, they will contribute to requirement-setting and benefit from enhanced safety and better-planned public transport.

Tools - until now two tools has been selected for the implementation of the FAITH AI_TAF: 1) Spiderisk[17], a cyber security risk management toolkit created by the University of Southampton automating ISO27005 risk assessment via a knowledge-based approach[18]. FAITH will extend its knowledge base from cybersecurity to account for AI trustworthiness and risks to accuracy, privacy, rights, etc. for AI components. 2) AI Model Passport, a tool designed to help manage and track machine learning experiments by integrating Data Version Control (DVC) [19] and MLflow [20]. It simplifies the process of organizing AI project into different phases, tracking changes, and logging results. The current AI Model Passport [21] will be extended for a broad spectrum of AI/ML pipelines and services.

Testbed - identification of the installation area on board the train and design of the on-board box to fit in the dedicated space; preparation of connections and communication interfaces towards Trenitalia trains; connections with the certification authorities for the certification of on-board cabin regarding safety, security, electromagnetic and electric compliance. All the Testbed activities have been carried out by the project partner Mermec Engineering [22].

5. Conclusions

In this paper, we presented the initiative undertaken within the FAITH project through the definition of a framework and appropriate metrics to calibrate the functional and non-functional characteristics of AI systems with the ultimate goal of achieving trustworthiness. This is obtained through a holistic and comprehensive risk management framework for AI and with extensive demonstration activities. Among the seven pilots planned in FAITH, in this paper, we focused on the transportation pilot, detailing some of its technological features that rely on pervasive vision systems, deep learning models, and behavioural analysis, describing specific features designed to achieve privacy by design. We then outlined the phases of this pilot, showcasing some preliminary activities and identifying a roadmap for future development.

Declaration on Generative AI

During the preparation of this work, the authors Grammarly in order to: grammar and spelling check, paraphrase and reword. After using this tool, the authors reviewed the text as needed and take full responsibility for the publication's content.

References

[1] Fostering Artificial Intelligence Trust for Humans towards the optimization of trustworthiness through large-scale pilots in critical domains (FAITH), 2024. URL: https://faith-ec-project.eu/, Last retrieved on 19 5 2025.

[2] A. Jobin, M. Ienca, E. Vayena, The global landscape of AI ethics guidelines, Nature Machine Intelligence 1 (2019) 389–399. doi:10.1038/s42256-019-0088-2.

[3] High-Level Expert Group on Artificial Intelligence, 2018. URL: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai, Last retrieved on 19 5 2025.

[4] EU, Regulation (EU) 2024/1689 of the EU parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (EC) no 300/2008, (EU) no 167/2013, (EU) no 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and

directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (artificial intelligence act), 2024. URL: http://data.europa.eu/eli/reg/2024/1689/oj.

[5] OSTP, Blueprint for an AI bill of rights. Making automated systems work for the American people, 2022. URL: https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/.

[6] NIST, Artificial intelligence risk management framework (AI RMF 1.0), 2023. URL: https://www.nist.gov/itl/ai-risk-management-framework, Last retrieved on 19 5 2025.

[7] ENISA, Cybersecurity of AI and standardisation, 2023. URL: https://www.enisa.europa.eu/publications/cybersecurity-of-ai-and-standardisation, Last retrieved on 19 5 2025.

[8] ISO/IEC TS 5723:2022 Trustworthiness—vocabulary, 2022. URL: https://www.iso.org/obp/ui/en/#iso:std:81608:en, Last retrieved on 19 5 2025.

[9] NIST, Engineering trustworthy secure systems, 2022. URL: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-160v1r1.pdf, Last retrieved October 27, 2025.

[10] G. P. Fettweis, P. Grünberg, T. Hentschel, S. Köpsell, Conceptualizing trustworthiness and trust in communications, arXiv preprint (2024). doi:10.48550/arXiv.2408.01447.

[11] E. Rama, M. Ayache, R. Buchty, B. Bauer, M. Korb, M. Berekovic, S. Mulhem, Trustworthy integrated circuits: From safety to security and beyond, IEEE Access (2024). doi:10.1109/ACCESS.2024.3400685.

[12] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities, MIT Press, 2023. URL: https://fairmlbook.org/, Last retrieved on 19.0192025.

[13] C. Sanderson, D. Douglas, Q. Lu, Implementing responsible ai: Tensions and trade-offs between ethics aspects, in: 2023 International Joint Conference on Neural Networks (IJCNN), IEEE, 2023, pp. 1–7. doi:10.1109/IJCNN54540.2023.10191274.

[14] B. Chander, C. John, L. Warrier, K. Gopalakrishnan, Toward trustworthy artificial intelligence (tai) in the context of explainability and robustness, ACM Computing Surveys (2024). doi:10.1145/3675392.

[15] G. R. Leone, A. Carboni, S. Nardi, D. Moroni, Toward pervasive computer vision for intelligent transport system, in: 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), IEEE, 2022, pp. 26–29. doi:10.1109/PerComWorkshops53856.2022.9767376.

[16] A. Carboni, G. R. Leone, S. Nardi, A. Corrado, D. Moroni, A novel smart camera network for real time public transport monitoring and surveillance, in: 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2023, pp. 5223–5228. doi:10.1109/ITSC57777.2023.10421861.

[17] University of Southampton, Spyderisk, 2024. URL: https://github.com/Spyderisk.

[18] System Security Modeller, 2022. URL: https://zenodo.org/record/6676219.

[19] Data Version Control, website. URL: https://dvc.org, Last retrieved on 19 5 2025.

[20] MLflow, website. URL: https://mlflow.org, Last retrieved on 19 5 2025.

[21] S. Colantonio, A. Berti, R. Buongiorno, G. Del Corso, E. Pachetti, M. A. Pascali, C. Kalantzopoulos, V. Kalokyri, H. Kondylakis, N. Tachos, D. Fotiadis, V. Giannini, S. Mazzetti, D. Regge, N. Papanikolaou, K. Marias, M. Tsiknakis, Ai trustworthiness in prostate cancer imaging: a look at algorithmic and system transparency*, in: 2023 IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biology, 2023, pp. 79–80. doi:10.1109/IEEECONF58974.2023.10404432.

[22] Mermec Engineering, website. URL: https://mermec-engineering.com/en/.