30 APRIL 2025

NEWSLETTER 4/2025

TALON PROJECT

TALON

Autonomous and self-organised artificial intelligent orchestrator for a greener industry 5.0

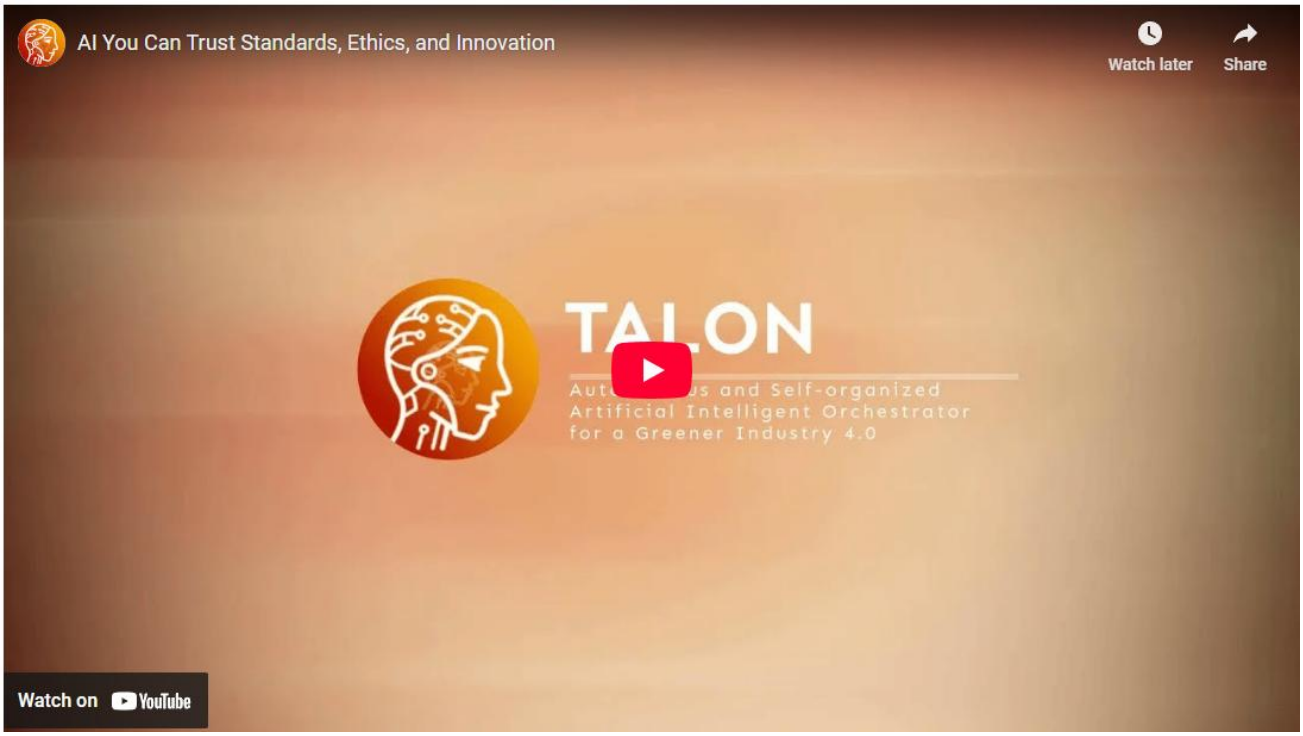
talon-project.eu

EDITORIAL

This newsletter presents a summary of the primary dissemination outcomes of the project, which include organization of events and technological developments. Specifically, one (1) publication, one (1) webinar, and one (1) blog post with the results of the TALON webinar.

Stylianos Trevlakis, InnoCube

AI YOU CAN TRUST: STANDARDS, ETHICS, AND INNOVATION



This webinar, initiated and hosted by the Talon Project, explores approaches and standards for building trustworthy AI, a complex challenge given AI's increasing sophistication in models, applications, and infrastructure management.

John Soldatos from Netcompany Intrasoft welcomed attendees to the webinar, "AI You Can Trust: Standards, Ethics, and Innovation," hosted by the Talon Project. He explained the webinar's aim to illuminate approaches and standards for building trustworthy AI, highlighting the increasing complexity of AI's sophistication in models and applications, including infrastructure management. He then introduced the four speakers: Christos Emanuilidis (University of Groningen), Leon Limonad (IBM Research), Renetta Polemi (UBITECH), and Renetta Polemi (Tastilio), emphasizing their expertise in industrial aspects, standardization, and the ethical considerations of trustworthy AI. Soldatos concluded by stating he would proceed directly to the first speaker.

Trust vs. Trustworthiness in AI

Christos Emanuilidis from the University of Groningen highlighted the distinction between trust and trustworthiness in AI. While explainable AI can increase trust, it doesn't guarantee trustworthiness.

- Trustworthiness encompasses reliability, resilience, robustness, safety, transparency, usability, and controllability.

- It's about verifiable ability to meet expectations.
- He emphasized the importance of human-AI teaming in decision-making, where agency can shift between humans and AI.
- Standards play a crucial role in achieving trustworthiness, offering benefits like improved service quality, growth, competitive edge, and regulatory compliance.
- The Humane project, which employs various learning paradigms like active learning and swarm learning, aims to demonstrate trustworthy AI in diverse use cases.

Situation-Aware Explainability (SACS)

Lior Limonad from IBM Research presented SACS, a framework for generating meaningful explanations about business processes.

- SACS analyzes process event logs to generate process, causal, and explainable AI (XAI) views.
- These views, combined with user inquiries, are used by a large language model (LLM) to create tailored explanations.
- A dedicated evaluation scale, incorporating fidelity, interpretability, trust, and curiosity, assesses the quality of these explanations.
- A user study revealed that adding knowledge improves fidelity but can reduce interpretability.
- Another study demonstrated the scale's usefulness in comparing different LLMs for generating explanations.
- The SACS library is open-source, encouraging community feedback and improvement.

Talon's Approach to Trustworthy AI

Sofia Karagiorgou from UBITEC discussed Talon's approach to trustworthy AI.

- The project focuses on an AI orchestrator, security constraints via blockchain, and digital twins for explainability.
- Challenges include the dynamic nature of AI ethics, the black box nature of AI models, and the context-dependence of ethical considerations.
- Talon's breakthroughs include an AI theoretical model for dimensioning hardware, a multimodal data ops and MLOps pipeline, and a zero-touch smart orchestrator for AI model serving.
- The project emphasizes a continuous journey towards trust in AI, with ethical considerations as a guiding compass.

Challenges and Efforts in AI Trustworthiness

Renetta Polemi from Trastilio addressed challenges in AI trustworthiness, particularly concerning the EU AI Act.

- She highlighted the need for harmonized standards, robust risk management frameworks that consider social and ethical threats, and standardized trustworthiness schemes for auditors.

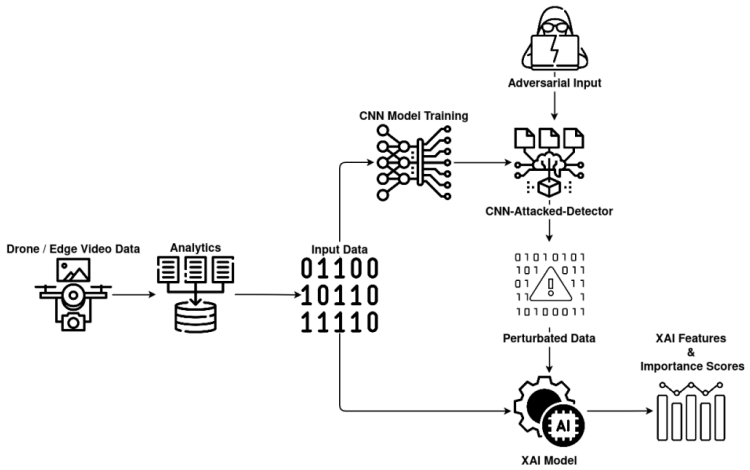
- The lack of AI certification authorities and the need for training on AI trustworthiness were also emphasized.
- Projects like FAITH are developing trustworthiness risk management frameworks and tools, incorporating human factors and iterative processes.

Conclusion

The webinar underscored the importance of moving beyond trust to trustworthiness in AI. Key takeaways include the need for harmonized standards, robust risk management frameworks, explainability solutions, and ongoing research into AI trustworthiness assessment. The discussion also highlighted the challenges of explaining complex AI models like LLMs and the need for multidisciplinary collaboration to address the ethical and societal implications of AI. The open-source nature of projects like SACS and the ongoing development of frameworks like FAITH offer promising paths towards building AI systems that are not only powerful but also trustworthy.

ADVERSARIAL EXPLANATIONS FOR INFORMED CIVILIAN AND ENVIRONMENTAL PROTECTION

Abstract: Combating crime and conditions of high physical risk in cities, the environment, and critical infrastructures requires a multifaceted approach. For sensitive problems, such as advanced situational awareness in the fields of civilian applications and environmental protection, Artificial Intelligence (AI) and Neural Network (NN) adoption has been slow due to concerns about their reliability, leading to several algorithms for explaining their decisions. Despite the possibilities for AI in critical infrastructure protection and civilian applications, many challenges still exist. For instance: (i) there are complex and high risks meaning that AI systems need to be transparent and interpretable to gain decision-maker trust; (ii) AI models may be vulnerable to imperceptible manipulations of input data even without any knowledge about the AI technique that is used; (iii) the need to efficiently process distributed, multimodal and big data coming from different, but however cheap, Internet of Things (IoT) and sensory devices (e.g., drones, cameras, accelerometers, telemetry, geomagnetic field, and proximity sensors); and (iv) many AI methods based on Machine Learning (ML) require huge amounts of training data, resulting in a Big Data computation problem. We introduce, benchmark, and demonstrate an adversarial explanations approach that we can efficiently tackle both adversarial robustness and explanation complexity of AI systems. To achieve this, we train robustified NNs and transparent explainers on big imagery data and leverage the attacks' knowledge as explanations to gain greater fidelity to the AI model. The merit of the proposed approach is that the new and robustified model has a great performance against new, unseen types of perturbations and attacks. This way, we pave the adoption of more informed and responsible AI integration in sensitive application domains.



T. Anastasiou, I. Pastellas and S. Karagiorgou, "Adversarial Explanations for Informed Civilian and Environmental Protection," 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 2024, pp. 2672-2681, doi: 10.1109/BigData62323.2024.10825734.

TALON

ENGINEERING
THE DIGITAL TRANSFORMATION COMPANY

ERICSSON

MINDS

netcompany
intrasoft

INTERNATIONAL
HELLENIC
UNIVERSITY

Kingston
University
London

CERTH
CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

uni.lu
UNIVERSITÉ DU
LUXEMBOURG

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

EIGHTBELLS
Independent Research & Consultancy

UBITECH
ubiquitous solutions

SIDROCO

InnoCube

factor
Ingeniería y Decoletaje S.L.

EXOS
Operational Consulting

PROBOTEK