



The AI Data Robotics  
Association

# Future-Ready

On-Demand Solutions with  
AI, Data, and Robotics

Supported by AI4Europe + Adra-e



Funded by  
the European Union

# Trustworthy AI: main innovations and future challenges

## Objectives:

- share the current and promising results of the sibling projects about trustworthy AI
- propose new challenges to be addressed for a better acceptance of AI solutions by the industrial domain

## Agenda:

1. Introduction (5') – M. Barreteau
2. Project pitch about 1-3 remarkable results + 1-3 future challenges (around 10' per project)
3. Panel discussion (around 30'): Q&A (between speakers and audience)
4. Final selection of the main results and challenges (15') to prepare the final sheet

# TrustWorthy AI Cluster

Working towards a Trustworthy AI

#TrustWorthyAICluster

AUTOFAIR

enexa

EVENFLOW  
Robust Learning and Reasoning for Complex Event Forecasting



SAFEXPLAIN

SustainML

TALON

Tuples  
TRUSTWORTHY AI

ULTIMATE

Horizon Europe supports **nine initiatives** to boost solid and trustworthy Artificial Intelligence across Europe

Supported by:



Human-compatible AI with guarantees, the Horizon Europe project (“AutoFair”), seeks to address needs for trusted AI and user-in-the-loop tools and systems in a range of industry applications.

## AUTOFAIR

EVENFLOW aims to develop hybrid learning techniques for complex event forecasting, which combine deep learning with logic-based learning and reasoning into neuro-symbolic forecasting models.



The SAFEXPLAIN project is laying the foundation for Critical Autonomous AI-based Systems (cars, trains, satellites) that are smarter and safer by ensuring that they adhere to functional safety requirements. It ensures that the Deep Learning (DL) technology needed to support most future advanced software functions in CAIS is explainable and safe and provides a novel and flexible approach to the certification – and hence adoption – of DL-based solutions, thus closing the gap between FUSA requirements and the black-box nature of DL solutions.



TALON introduces an AI orchestrator that envisions transforming the I5.0 into an automated intelligent platform by exploiting advances in edge networks and **bringing intelligence near sensors** in embedded systems with limited computational, storage, and communication resources, as well as the integration of advanced and adaptive sensors and perception.



ULTIMATE will pioneer the development of industrial-grade hybrid AI based on three stages to ensure trustworthiness and promote the widespread adoption of hybrid AI in industry.



ENEXA builds upon novel and promising results in knowledge representation and machine learning to develop scalable, transparent, and explainable hybrid machine learning algorithms that combine symbolic and sub-symbolic learning.



The REXASI-PRO project aims to release a novel engineering framework to develop green and trustworthy Artificial Swarm Intelligence solutions for the collaboration cooperation between a swarm of autonomous wheelchairs and flying robots to help people with reduced mobility.



This project envisions a sustainable, interactive ML framework development for Green AI that will comprehensively prioritize and advocate energy efficiency across the entire life cycle of an application and avoid AI-waste.



The project will take a more integrated and human-centered approach to the development of P&S tools, in order to increase confidence in these systems and accelerate their adoption. In particular, we will design methods that combine the power of data-driven and knowledge-based symbolic AI.

## TrustWorthy AI Cluster

Working towards a Trustworthy AI

#TrustWorthyAICluster



# Index

## Session 1

### Verifiable Robustness, Energy Efficiency, and Transparency for Trustworthy AI

#### ULTIMATE

Trustworthiness for the development of industrial-grade hybrid AI

Michel BARRETEAU (Thales cortAIx Labs)

#### EVENFLOW

Hybrid learning techniques for complex event forecasting

Nikos KATZOURIS (Demokritos)

#### TALON

AI orchestrator to bring intelligence near sensors

Stylianos TREVLAKIS (InnoCube)

#### SAFEXPLAIN

Smarter and safer critical autonomous AI-based systems

Jaume ABELLA (BSC)

Supported by:



Funded by  
the European Union

# Jaume ABELLA

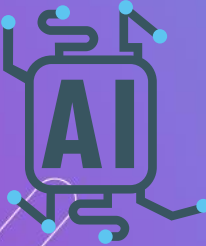
**HIGH PERFORMANCE EMBEDDED SYSTEMS LABORATORY  
DIRECTOR (BSC)**

**SAFEXPLAIN**

## **Bio:**

Dr. Jaume Abella co-leads the CAOS research group (embedded systems group) at BSC, where he has worked since 2009. He received his PhD in 2005 and worked as Senior Researcher at Intel Corporation from 2005-2009. He is the **coordinator of the SAFEXPLAIN project** and has been the PI at BSC of several EU projects on enabling the use of high-performance hardware and software, as well as **AI in safety-critical systems**, in different funding frameworks such as Horizon Europe (NimbleAI, SAFEXPLAIN), Chips JU (SMARTY, ISOLDE, REBECCA), H2020 (SELENE, DeRISC, SAFURE), ECSEL (FRACTAL) and ARTEMIS (VeTeSS).

Dr. Abella holds **15 patents issued**, has **published around 250 papers** in top peer-reviewed conferences and journals, has co-advised 15 PhDs and 20 Master theses, and co-founded a successful spinoff providing software services in avionics and automotive (Maspatechnologies S.L., now Rapita Systems S.L.).



Supported by:



Funded by  
the European Union

Theme

04



# SAFEXPLAIN

The SAFEXPLAIN project is **laying the foundation for Critical Autonomous AI-based Systems** (cars, trains, satellites) that are smarter and safer by **ensuring that they adhere to functional safety requirements**. It ensures that the **Deep Learning (DL) technology** needed to support most future advanced software functions in AI-based critical systems **is explainable and safe** and provides a novel and flexible approach to the **certification** – and hence adoption – of DL-based solutions, thus closing the gap between functional safety requirements and the black-box nature of DL solutions.

Supported by:



Funded by  
the European Union

# SAFEXPLAIN



[www.safexplain.eu](http://www.safexplain.eu)

Safe and explainable critical  
embedded systems based on AI

Jaume Abella (project coordinator)



Supported by:



This Project has received funding from the European Union's Horizon Europe programme under grant agreement number 101069595



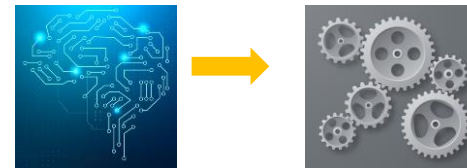
# Critical Embedded Systems and AI

- **AI techniques** are at the very heart of the realization of **advanced software functions** such as **computer vision for object detection** and tracking, path planning, driver-monitoring systems,...
- **Till today, AI mostly in fail-safe systems** (i.e. systems with a safe state)
  - Supervise the AI-based functionality and, if it fails, switch it off, switch to a degraded mode, etc.
  - AI software without safety requirements
- **No longer possible with fully-autonomous systems**
  - AI software inherits safety requirements, so it must adhere to a development process guaranteeing functional safety
- Challenge: AI software development does not adhere to the process for usual control software
  - **Empirical** development without explicit and precise requirements
  - **Data-based** software design
  - **Not “correct-by-construction”**
  - **Monolithic** architecture

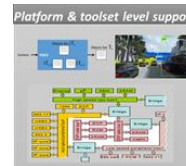
# SAFEXPLAIN ambition

- Architecting DL solutions **enabling certification/qualification**

- Making them **adhere to “safety culture”**



- Preserving **high performance**




- Tailoring solutions to **varying safety requirements** (e.g., different safety needs for a coffee machine and a plane)



**SAFEXPLAIN in very good shape to meet its objectives**

OCT 2022 – SEP 2025



Safe and Explainable  
Critical Embedded Systems based on AI

---

BARCELONA SUPERCOMPUTING CENTER (BSC)  
<https://www.bsc.es/>

IKERLAN, S. Coop (IKR)  
<https://www.ikerlan.es/>

AIKO SRL (AIKO)  
<https://www.aikospace.com/>

RISE RESEARCH INSTITUTES OF SWEDEN AB (RISE)  
<https://www.ri.se/>

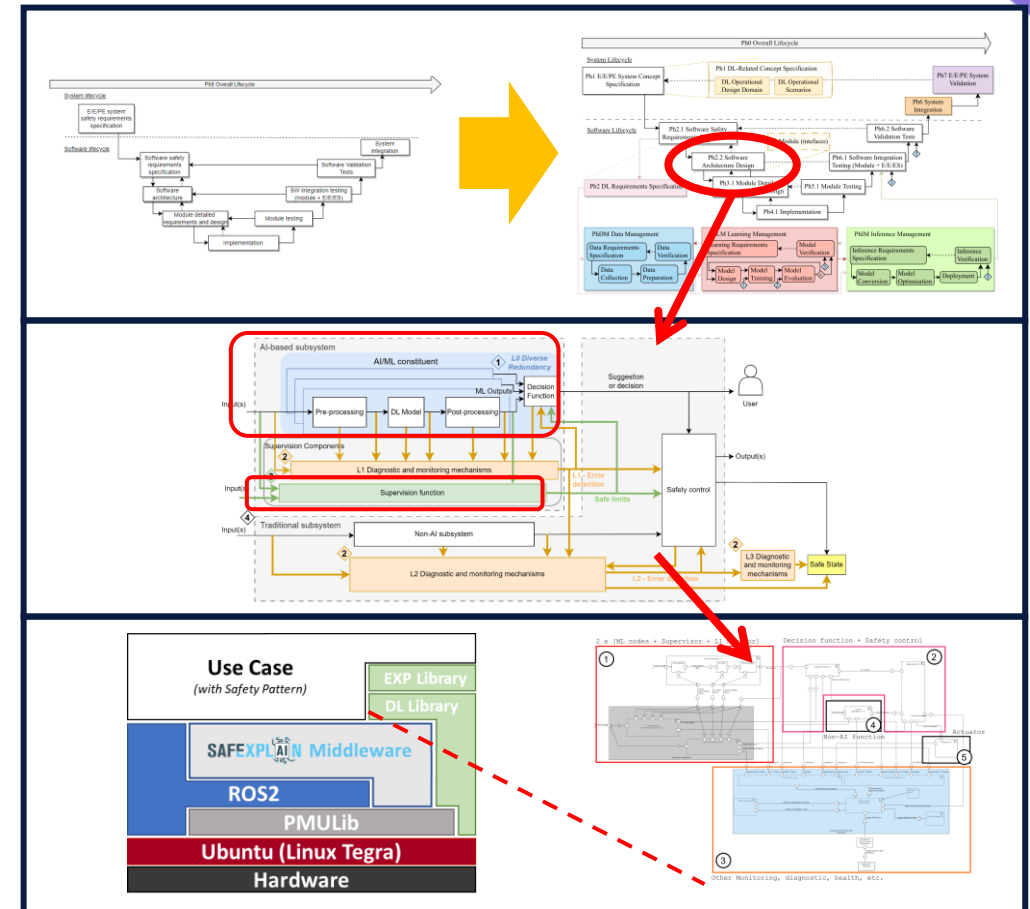
NAVINFO EUROPE BV (NAV)  
<https://www.navinfo.eu/>

EXIDA DEVELOPMENT SRL (EXI)  
<https://www.exida-eu.com/>

# Remarkable Results

- Enable the use of AI inheriting safety requirements

- Extended **development process**
  - Map safety requirements to AI components
  - Set compliance steps in accordance with standards
- New **software architecture**
  - Build an architecture with appropriate redundancy, monitoring and supervision
  - Preserve safety despite AI model limitations
- Suitable **software stack**
  - Modular, efficient, and providing all safety-related services needed



# Future Research/Innovation Challenges

- Concepts **passed TÜV Rheinland review** against certification/qualification acceptable practices
- So far technologies prototyped and evaluated in **industrial demonstrators**
  - Automotive, space and railway domains
- But still there is a **gap for true commercial adoption**
  - Run **pilot projects** with commercial-size use cases
  - Adjust and **automate** processes and tools
  - **Mature** the toolset and, if needed, qualify it
- (longer term) Extend the solution to **allow newer AI solutions**
  - Systems learning during operation
  - Use of LLMs, genAI, etc.

# Nikos KATZOURIS

## Researcher (Demokritos)

EVENTFLOW technical manager

## Bio:

Nikos Katzouris is a researcher at the institute of Informatics & Telecommunications at NCSR Demokritos. His research focuses on combinations of Machine Learning with Knowledge Representation and Reasoning. In particular, techniques that can improve the transparency, robustness and trustworthiness of learning-based AI systems, with a special focus on temporal, event-based applications.



Supported by:



The AI Data Robotics Association



Funded by  
the European Union

Theme

02

# EVENFLOW

EVENFLOW aims to develop hybrid learning techniques for complex event forecasting, which combine deep learning with logic-based learning and reasoning into neuro-symbolic forecasting models.

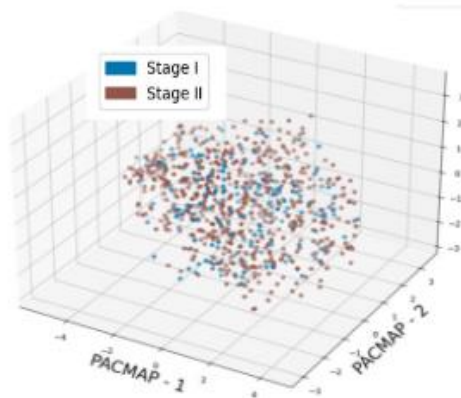
Supported by:



Funded by  
the European Union

# EVENFLOW

EVENFLOW develops scalable, formally verifiable and explainable neuro-symbolic techniques for detecting imminent critical situations before they actually occur.



## Personalized Medicine

Gene expression data → VAE → virtual patients →  
patient trajectories →  
temporal patterns of disease progression →  
condition deterioration forecasts



## Industry 4.0

AGV robots' perception + execution plans +  
robots' mobility patterns →  
event forecasting model →  
decentralized collision, or delay forecasts



## Infrastructure maintenance

Water pipe accelerometer data + water pipe  
network spatial knowledge →  
localized leakage predictions.

Supported by:

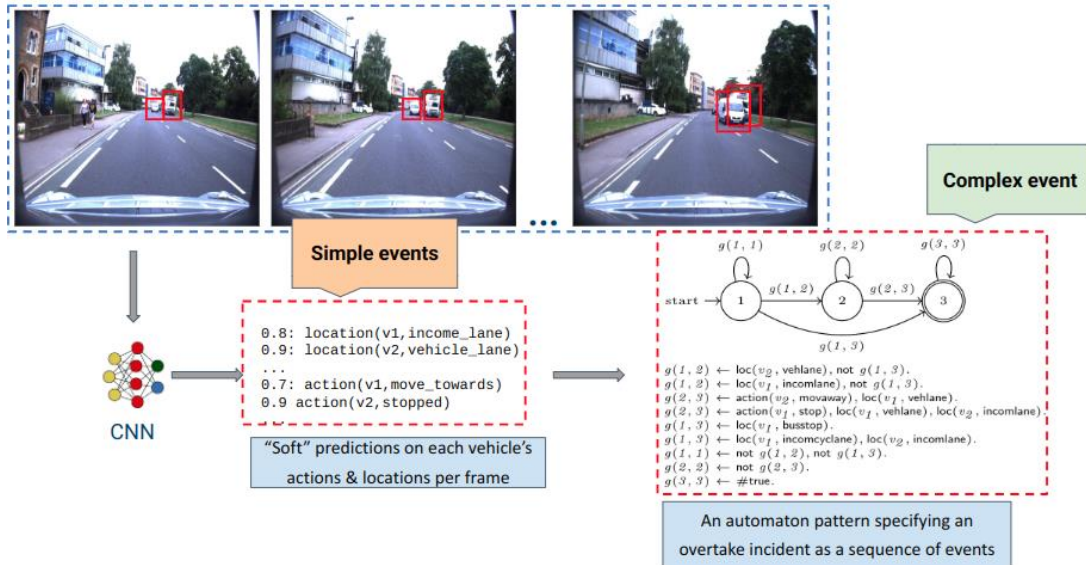


Funded by  
the European Union

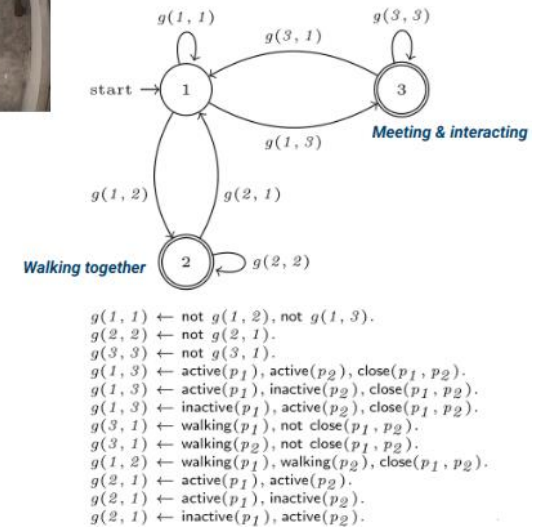


# 1. Temporal Neuro-Symbolic Learning & Reasoning

Overtake incident



Symbolic model used with NeSy methods:



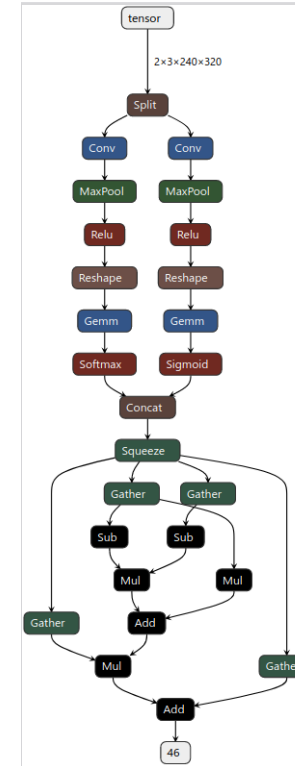
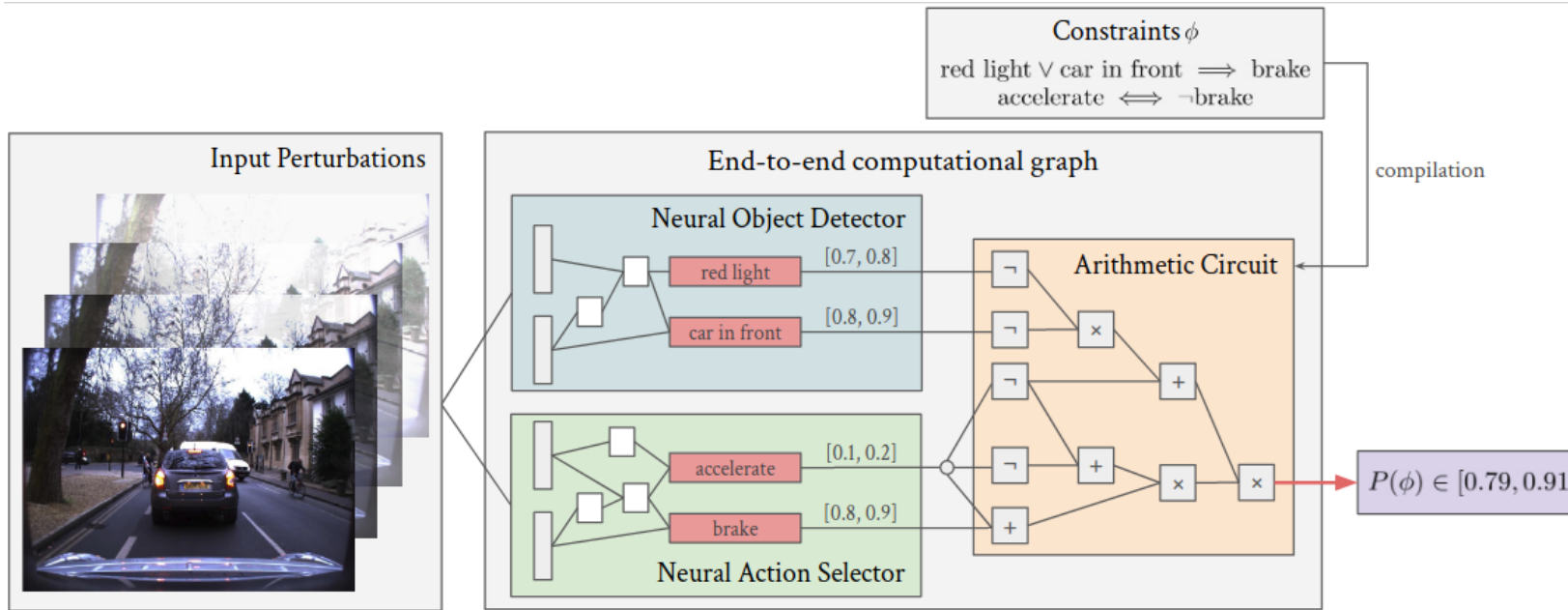
Model	Training time (min)	micro-F1 (test)	#params
CNN+LSTM stack	7	0.54	7897603
<b>CNN+NeSy-Automata (new method)</b>	8	<b>0.94</b>	2520580
DeepProbLog	Out of memory*		
DeepStochLog	Out of memory*		
NeurASP	Out of memory*		

\*128 GB

- End-to-end training of NNs with event patterns represented in temporal logics & automata.
- Scalable, data efficient, explainable.
- Much better generalization in out-of-distribution settings.
- Joint neural training and event pattern learning.

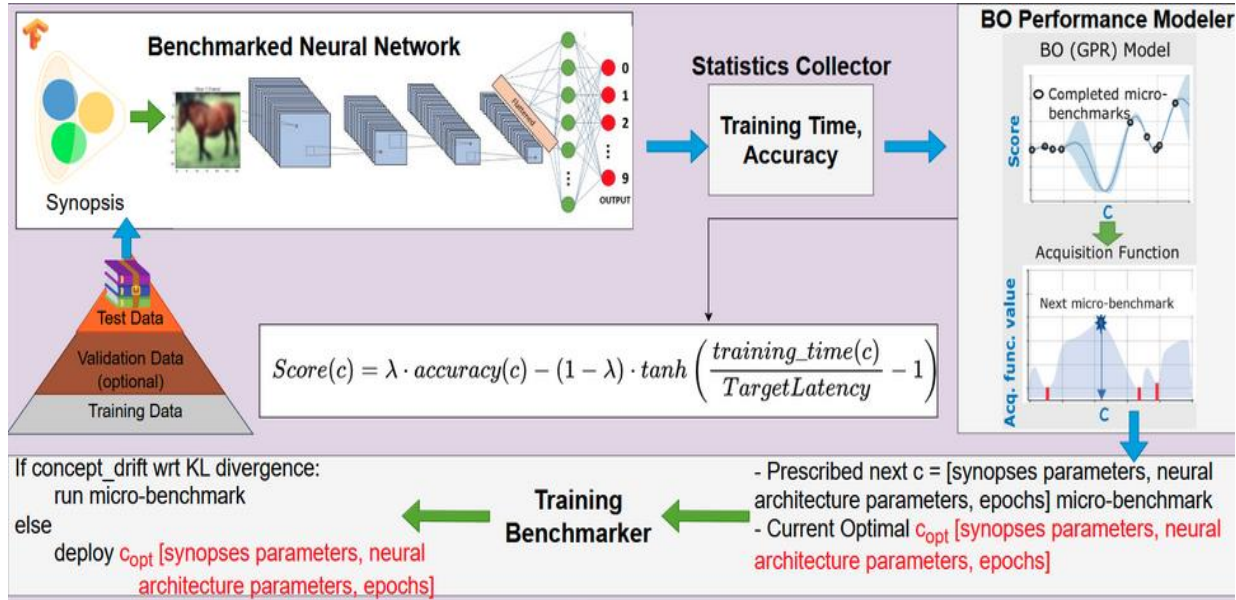


# 2. Scalable Neuro-Symbolic Verification



- Formally verify the robustness of hybrid systems, consisting of neural & symbolic components.
- Scalable, relaxation-based approach, combining NN verification techniques with knowledge compilation.

# 3. Scalable Training



- Resource-efficient distributed training.
- Continuous learning of training time vs accuracy trade-offs in evolving data streams, fine-tuning in real time:
  - Employed neural architectures.
  - Size of ingested data via stream synopses-specific parameters.

# Future Challenges

- **Situational awareness for robust, transparent, self-adaptive AI systems.**
  - Perceive, comprehend, forecast, adapt.
- **Requirements/challenges:**
  - World "understanding".
  - Beyond language-based Generative AI (vision, time series, temporal evolution).
  - Structured knowledge, constraints & causal models can help:
    - Generalization/robustness, explainability, data efficiency.
  - **Neuro-symbolic world models.**
  - Formal guarantees in high-stakes applications?
  - Scalability & efficient collaboration in decentralized multi-agent settings?

# Stylianos Trevlakis

Director  
Researcher



Dissemination Manager



Stylianos E. Trevlakis was born in Thessaloniki, Greece, in 1991. He received the Diploma degree (5 year) in electrical and computer engineering (ECE) from the AUTH in 2016, and the Ph.D. degree from WCIP, AUTH in April 2022. Afterwards, he served with the Hellenic Army in the Research Office for nine months. From October 2017 until April, he was a Teaching Assistant with the Department of ECE, AUTH. Since 2022, he has been working with InnoCube as a Postdoctoral Researcher with focus on state-of-the-art research in conventional and AI-enabled wireless communication systems. His research interests lie in wireless communications, with emphasis on conventional and AI-enabled wireless communication systems, as well as communications and signal processing for biomedical engineering.



Supported by:



Theme

03

# TALON

TALON introduces an AI orchestrator that envisions transforming the I5.0 into an automated intelligent platform by exploiting advances in edge networks and bringing intelligence near sensors in embedded systems with limited computational, storage, and communication resources, as well as the integration of advanced and adaptive sensors and perception.

Supported by:



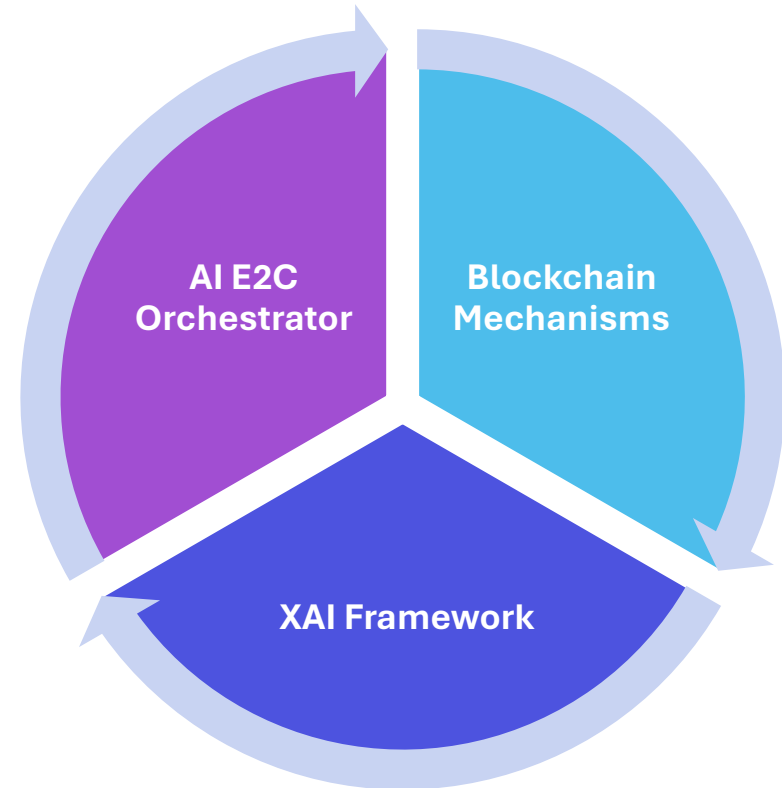
Funded by  
the European Union

# TALON's Remarkable Results

The TALON project has identified three remarkable results that are considered to have commercial and societal significance:

- Blockchain Mechanisms
- XAI Framework
- AI E2C Orchestrator

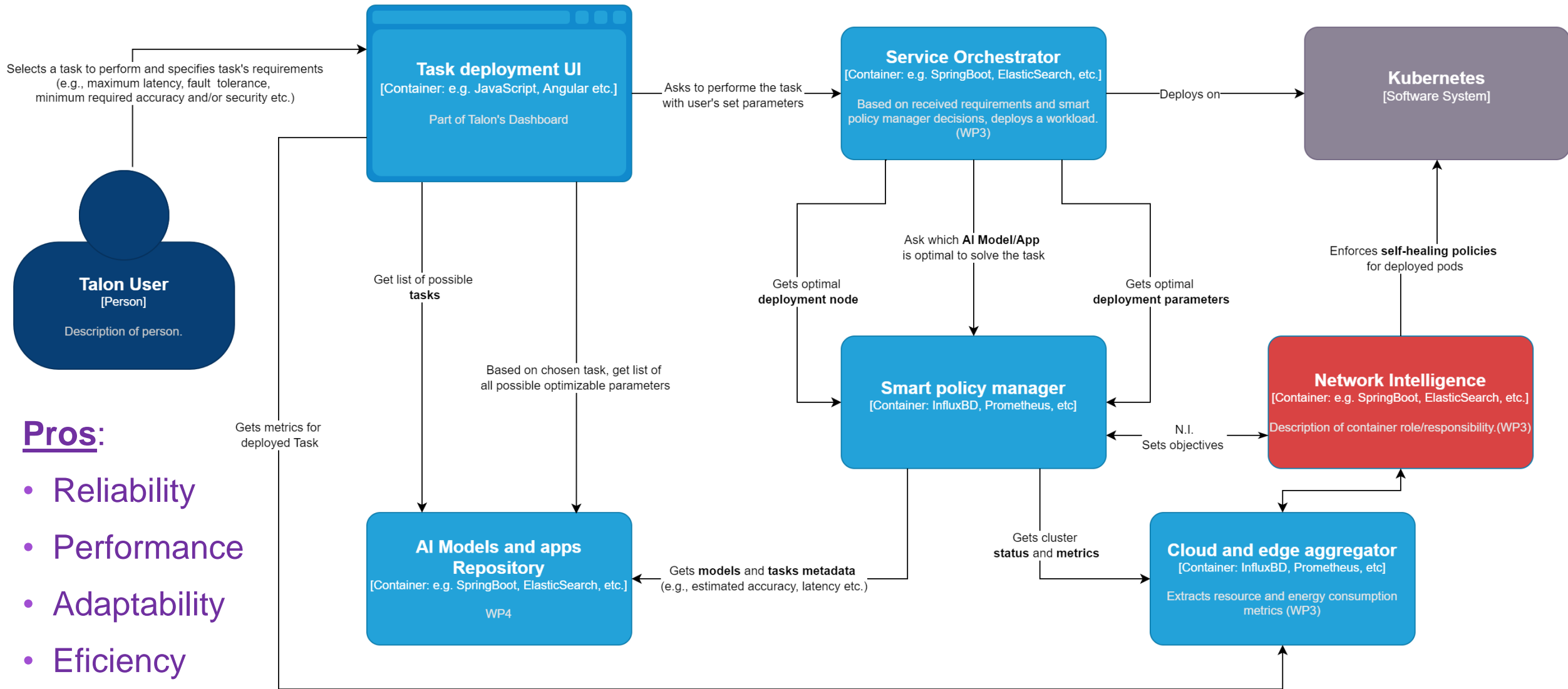
*Intelligence near the edge in a secure, explainable, and efficient manner!*



Supported by:



# TALON's AI E2C Orchestrator



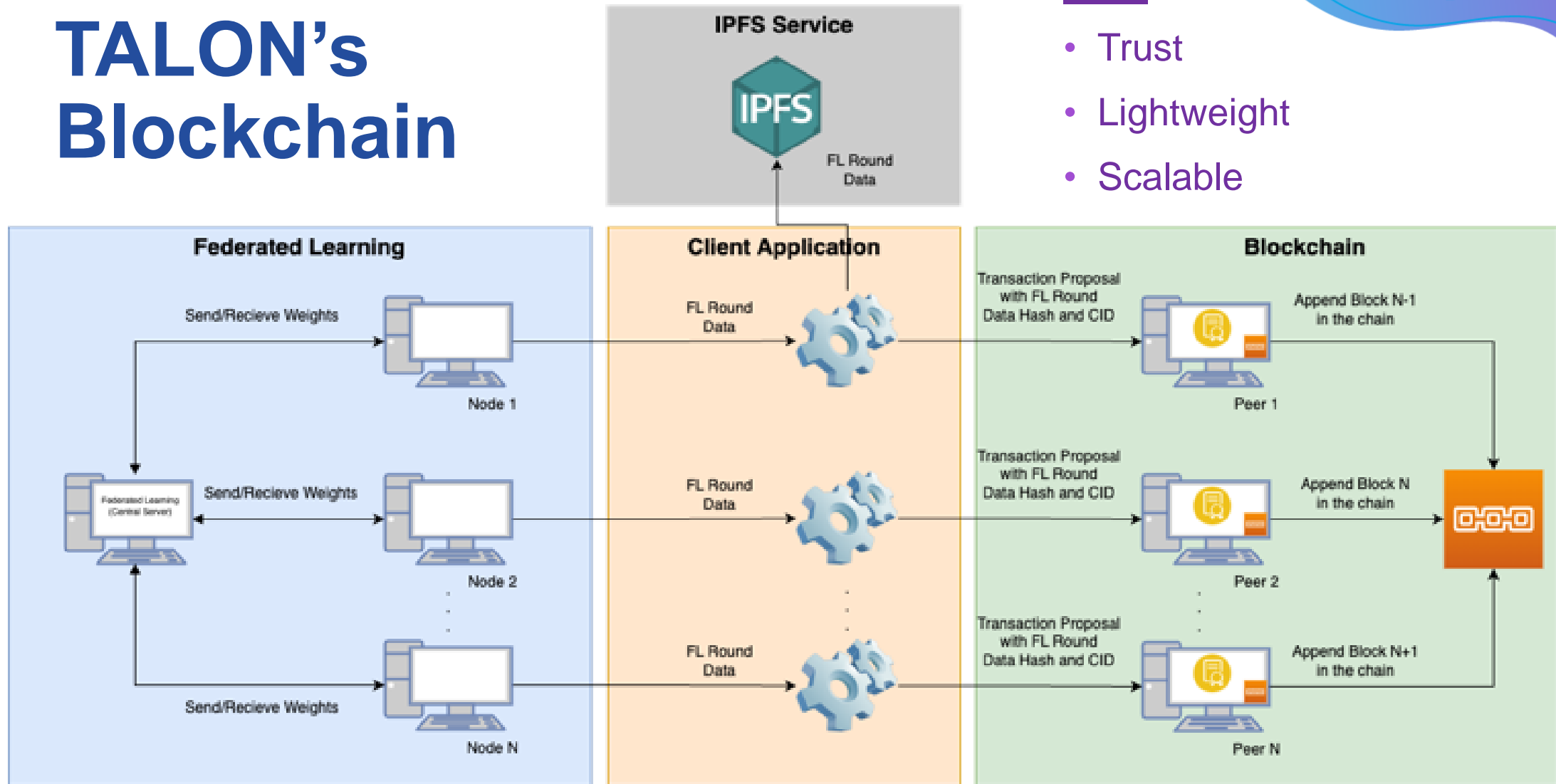
## Pros:

- Reliability
- Performance
- Adaptability
- Efficiency

Supported by:



# TALON's Blockchain



## Pros:

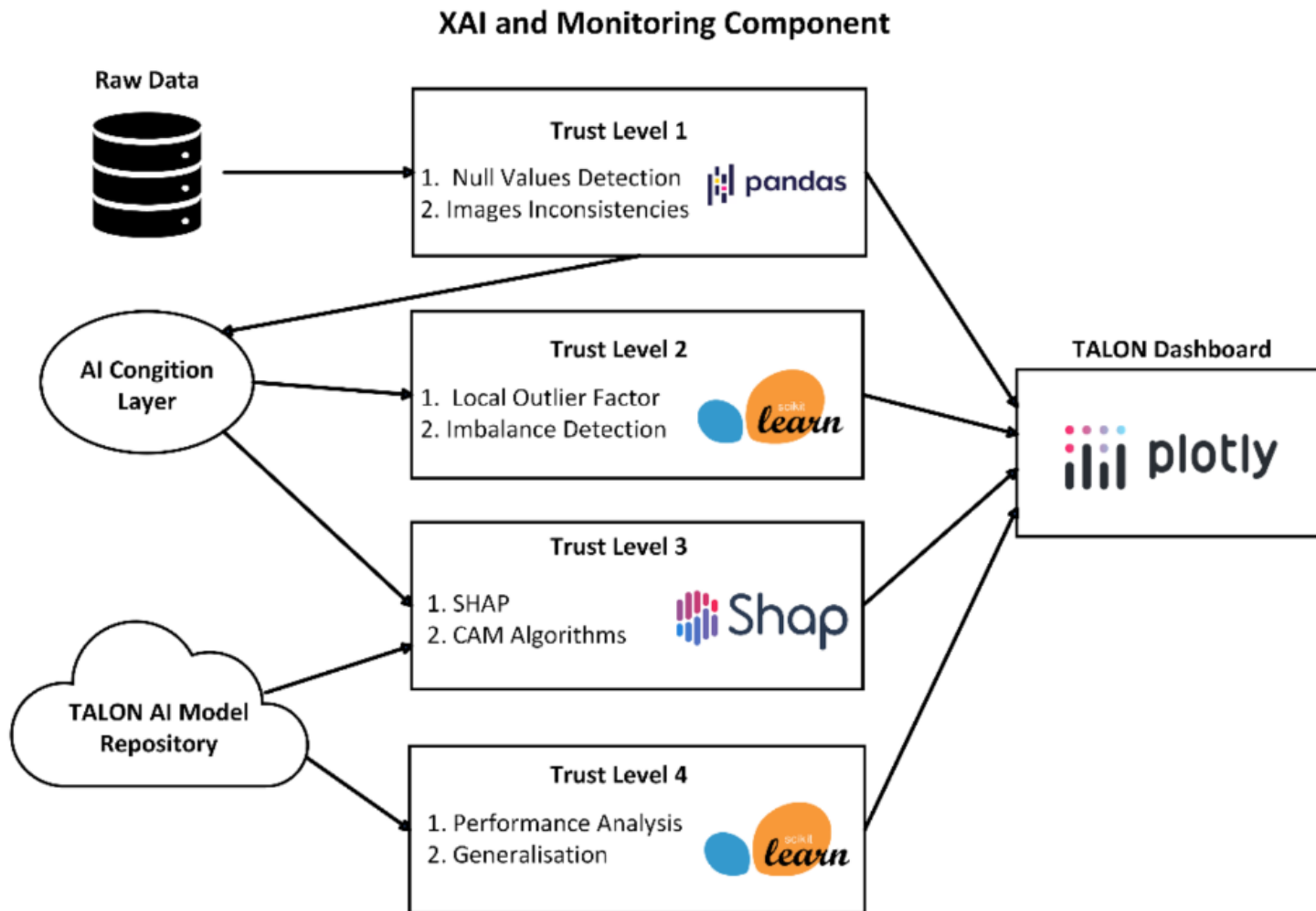
- Trust
- Lightweight
- Scalable

Supported by:





# TALON's Explainability Approach



## Motivation:

- XAI provides the necessary interpretability & transparency to the network's edge
- Enables HRC through interpretable and trustworthy AI procedures
- Attests for AI's decision-making in the industrial domain

## XAI Trust Levels (TrLs):

1. Raw data availability & quality
2. Data processing attestation & imbalance detection
3. Model agnostic explanations
4. Assess models robustness applicability

Supported by:



# TALON's Research/Innovation Challenges

- Scalability and Real-time Processing

Edge-to-cloud (E2C) & security



AI E2C orchestration to optimize the edge vs cloud resources.

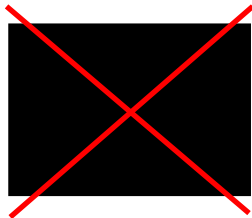


Lightweight hierarchical blockchain scheme.

- Scalability
- Real-time processing
- Security
- High processing power at the edge

- Explainability & Trustworthiness of AI Models

Black-box AI



Multi-level explanations, real-time interpretability & adaptability.



Advanced feedback for E2C AI model orchestration & optimization.

- Transparency
- Seamless integration
- Multi-level explanations

Supported by:



# Michel BARRETEAU

End-to-end Trustworthy AI expert (Thales cortAIx Labs France)

ULTIMATE coordinator

## Bio:

Dr. Michel Barreateau held his PhD (automatic parallelization) in Computer Science at Versailles University in 1997. Then he worked as an expert engineer on the iterative compilation for multimedia applications. He entered to Thales Research & Technology as a Research Engineer in 1999. Three years later, he was responsible of a tool devoted to the mapping of data-intensive signal and image processing applications onto parallel architectures. As a Programme Manager, he led many R&T projects during 15 years. He also managed an expertise working group about Algorithm Architecture Adequacy. Since 2018, he is in charge of transverse research activities on End-to-end Trustworthy AI. He mainly addresses the AI algorithm engineering challenges and related links with system and software levels for qualification or certification purpose of Thales critical AI-based systems. He is involved in several European civilian (e.g. as a coordinator of the ULTIMATE project devoted to Trustworthy hybrid AI) and Defence projects (incl. NATO) for Thales cortAIx Labs France.



Supported by:



Funded by  
the European Union

Theme

01

# ULTIMATE

ULTIMATE will pioneer the development of industrial-grade hybrid AI based on three stages to ensure trustworthiness and promote the widespread adoption of hybrid AI in industry

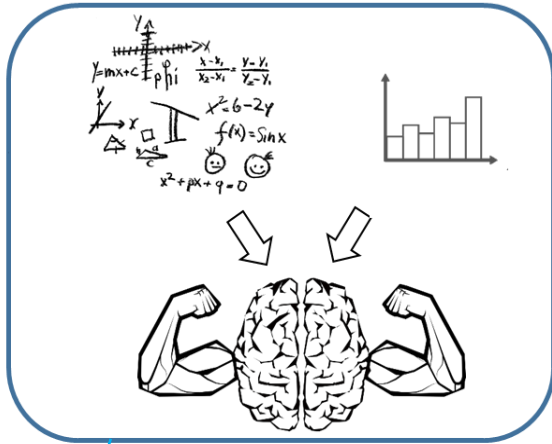
Supported by:



Funded by  
the European Union

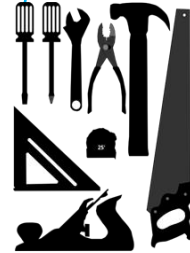
# ULTIMATE: 3 main stages to address trustworthiness regarding hybrid AI

Hybrid AI

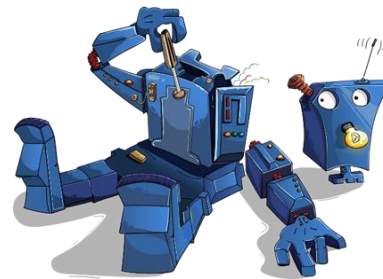


Develop **innovative architectures** to construct and train hybrid AI algorithms

**Design & Development**



Design **rigorous evaluation methodologies** with appropriate properties



**Evaluation**



Ensure the **ethical compliance**



**Trustworthiness**

Implement the hybrid AI algorithms under **operational conditions**

**Pre-industrialisation**



# Safe (autonomous) robots based on hybrid AI

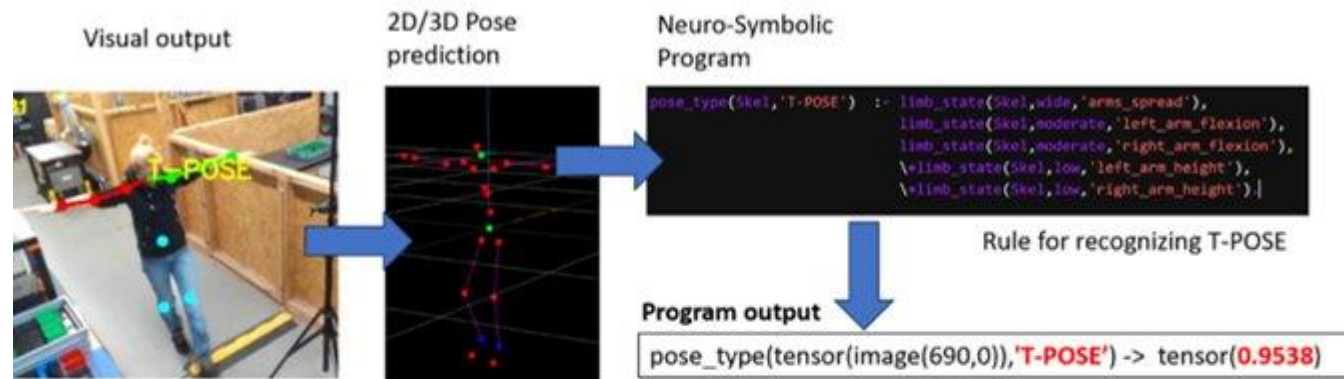
Develop innovative architectures to construct and train hybrid AI algorithms



Design & Development



- Context: area shared between humans and autonomous robots
- Hybrid AI: take benefit from the powerful connectionist AI (DL) and the explainable knowledge-based AI (e.g. rules), increase adaptivity to new situations



## Explanation:

```
limb_state(tensor(image(690,0)), wide, 'arms_spread') -> tensor(1.)
limb_state(tensor(image(690,0)), nobend, 'left_arm_flexion') -> tensor(1.)
limb_state(tensor(image(690,0)), nobend, 'right_arm_flexion') -> tensor(1)
limb_state(tensor(image(690,0)), moderate, 'left_arm_height') -> tensor(0.9700)
limb_state(tensor(image(690,0)), moderate, 'right_arm_height') -> tensor(0.9827)
```



Evaluation

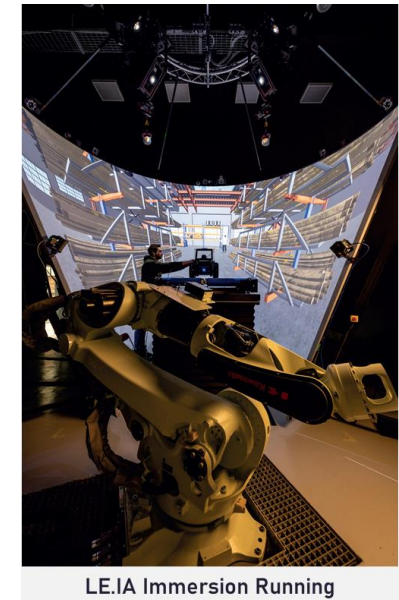
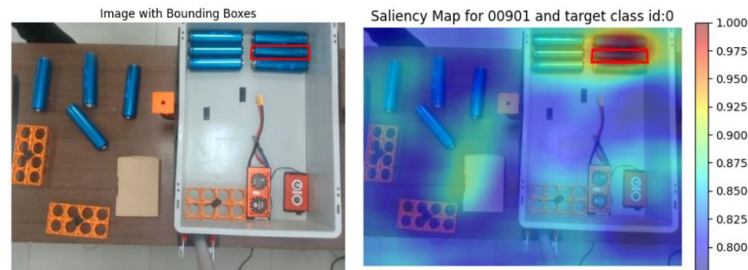


# Toolset for (hybrid) AI algorithm evaluation

- Statistical, empirical and formal evaluation toolset methods related to:
  - Reliability
    - Evaluation of confidence scores
    - Conformal anomaly detection
  - Robustness
    - Metamorphic robustness evaluation
    - Formal tools for robustness verification
    - Experimental robustness evaluation
    - Robustness evaluation based on domain adaptation
  - Explainability
    - Explanation evaluation
    - Visual explanation tools
- Open source or proprietary

Evaluation of 4 scores against 2,000 sets and 4 metrics:  
 AUC() / AUC(max\_fpr=0.2, kInd='convex\_hull') / MCC() / TPR(max\_fpr=0.05)

Scores Recorded layers	OoD 1: Left-out class	OoD 2: Uniformly random
Baseline()	0.803 / 0.380 / 0.508 / 0.241	0.587 / 0.169 / 0.457 / 0.880
KNN(k=31) (2, 25)	0.803 / 0.359 / 0.497 / 0.168	0.794 / 0.311 / 0.522 / 0.128
Energy()	0.832 / 0.436 / 0.542 / 0.278	0.749 / 0.243 / 0.589 / 0.016
GradNorm() (2, 30)	0.837 / 0.477 / 0.546 / 0.304	0.906 / 0.612 / 0.814 / 0.124





# End-to-end trustworthy AI (Value Sensitive Design based) methodology



- **Experimental** approach
- Addressing **both technical and ethical** trustworthy AI criteria by **key stakeholders** (end-user, AI developer, AI V&V) when implementing key engineering activities
- **Impacting the way stakeholders think, design, implement and evaluate** the AI solution aiming at achieving an **increased confidence** in AI for a successful implementation of AI based innovations

Information from D2.1 & D2.2										Use case (UC)				100%
WPS	Trustworthy Requirement	Category	Related Req.	Description	Comment	Priority	Considered in UC	Linked with developments of WPS?	Priority for WPS	Comments WPS	Yes/No/Maybe (Technical solution)	Yes/No/Maybe	Yes/No/Maybe	Please, explain your answer or how you are planning to tackle this.
TR_VSD_SPACE_27	Transparency	Traceability		Consider measures/mechanisms to ensure traceability (which input data was gathered and selected and how the result occurred)	Yes. The system should keep some sets of data corresponding to anomalies to enable further analysis or re-creating of the AI. [High]	HIGH	YES	YES	YES	PARTNER: All the results are gathered on log files for post-hoc analysis	YES	YES		
TR_VSD_SPACE_28	Transparency	Explainability		Consider to what extent the decision, and hence the outcome, made by the Hybrid AI can be understood. To what degree the system's decision influences the decision-making process? Ensure an explanation as to why the system took a certain choice resulting in a certain outcome that all the users can understand.	Yes. The ground team should be able to understand what rules of processes have led to the detection of an anomaly. So that the root-cause analysis is easier. And to better understand, when the system is likely to be correct or wrong so that the important decisions are taken correctly. [Medium]	MEDIUM	YES	NO	MAYBE	PARTNER: not considered for development but it will be needed some approximation for future since the developed technique is a deep learning approach used as black box	NO	MAYBE		
TR_VSD_SPACE_29	Transparency	Explainability		Consider procedures to analyse the data and to update this over time.	Yes. Further details needed	MEDIUM	YES	MAYBE	MAYBE	PARTNER: We need further details	MAYBE	MAYBE		
TR_VSD_SPACE_30	Transparency	Explainability		Consider whether you can examine retrospectively after the model training and development and whether you have access to the internal workflow of the model?	Yes. The external workflow of the model should be available to enable further analysis by an expert and determine if the AI rules are relevant enough. [Medium]	MEDIUM	YES	NO	MAYBE	PARTNER: We cannot provide mechanisms to make the model interpretable due to the architecture	NO	MAYBE		
TR_VSD_SPACE_31	Transparency	Communications		Implement measures/ mechanisms to inform the users the reasons and criteria behind the Hybrid AI outcomes.	Yes. When an anomaly is detected, the system should provide additional information (rules, meta-data, etc) to support the AI decision. [Medium]	MEDIUM	YES	YES	YES	PARTNER: We return the answer + confidence value provided by the net	YES	YES		
TR_VSD_SPACE_32	Transparency	Communications		Consider human psychology and potential limitations such as risk of confusion, confirmation bias or cognitive fatigue.	Yes. The additional information provided by the AI (see above) should be as complete as possible, so that the ground team analysis is as objective as possible. [Medium]	MEDIUM	YES	YES	NO	PARTNER: We return the answer + confidence value provided by the net	YES	NO		
TR_VSD_SPACE_33	Transparency	Communications		Consider clear communication on characteristics, limitations and potential shortcomings of the Hybrid AI (including potential bias)	Yes. The AI developer should deliver proper documentation, explaining the training procedures, and listing the limitations of the method, giving examples of potential future modes. [High]	HIGH	YES	YES	YES	PARTNER: Information provided in the user manual and deliverables	YES	YES		






# Future research / innovation challenges

1. Hybrid AI evaluation consolidated by knowledge  
Complete the evaluation of the connectionist AI part with the knowledge of the symbolic AI part
2. Trustworthy hybrid AI along the whole life cycle  
Extend the ULTIMATE results to cover all the stages of the hybrid AI life cycle in terms of trustworthy AI activities and governance, whatever the hybridization type
3. AI-based risk management environment  
Build a flexible and dedicated environment to manage AI risks for a specific critical domain based on an AI risk repository to fulfil the requirements of AI regulation document(s)

 humancompatible.org/  
 @AutoFair\_EU

**AUTOFAIR**

 evenflow-project.eu/  
 @EvenflowProject  
 EVENFLOW

**EVENFLOW**  
Robust Learning and Reasoning for Complex Event Forecasting

 safexplain.eu/  
 @SafexplainAI  
 SAFEXPLAIN

**SAFEXPLAIN**

 talon-project.eu/  
 @talon\_project  
 TALON

**TALON**

ultimate-project.eu/  
ULTIMATE

**ULTIMATE**

  
 enexa.eu/  
 @enexa\_eu



**TrustWorthy AI Cluster**

Working towards a Trustworthy AI

#TrustWorthyAICluster






  
**REXASI PRO**

rexasi-pro.spindoxlabs.com/  
  
  
 @REXASIPRO\_EU  
REXASI-PRO



  
 sustainml.eu/

**Tuples**  
TRUSTWORTHY AI  
 tuples.ai/  
 @tuplesAI  
 TUPLESAI



[humancompatible.org/](https://humancompatible.org/)

[enexa.eu/](https://enexa.eu/)

[evenflow-project.eu/](https://evenflow-project.eu/)

[rexasi-pro.spindoxlabs.com/](https://rexasi-pro.spindoxlabs.com/)

@AutoFair\_EU

@enexa\_eu

@EvenflowProject

@REXASIPRO\_EU

EVENFLOW

REXASI-PRO

**AUTOFAIR**



[safexplain.eu/](https://safexplain.eu/)

[sustainml.eu/](https://sustainml.eu/)

[talon-project.eu/](https://talon-project.eu/)

[tuples.ai/](https://tuples.ai/)

[ultimate-project.eu/](https://ultimate-project.eu/)

@SafexplainAI

@talon\_project

@tuplesAI

ULTIMATE

SAFEXPLAIN

TALON

TUPLESAI



Funded by the European Union

**TrustWorthy AI Cluster**

Working towards a Trustworthy AI

#TrustWorthyAICluster





The AI Data Robotics  
Association

# Thank you!

Supported by AI4Europe + Adra-e



Funded by  
the European Union