Assessing Trustworthiness on Hybrid AI Applications

On an ethical and legal overarching approach

Nuria Quintano Fernández

Tecnalia

Workshop “The way forward: future challenges on software engineering“

June 27th 2023, Milan

Agenda

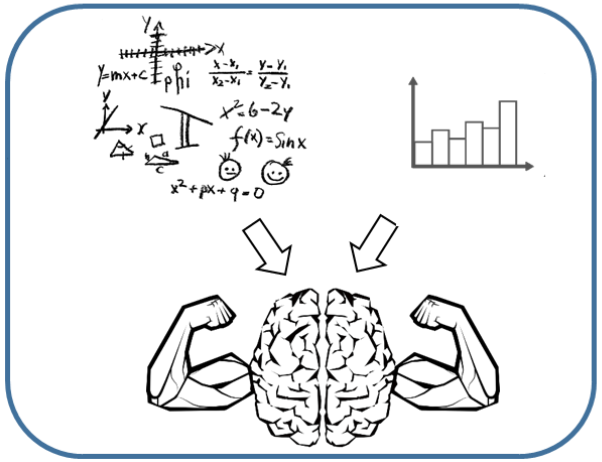
- ULTIMATE project
- Tecnalia's objectives in ULTIMATE
- Work already started: Hybrid AI trustworthiness requirements identification and prioritisation
- Future work: Hybrid AI trustworthiness requirements implementation, quantification and modelling
- References

ULTIMATE Project

- HORIZON-CL4-2021-HUMAN-01-01: **A HUMAN-CENTRED AND ETHICAL DEVELOPMENT OF DIGITAL AND INDUSTRIAL TECHNOLOGIES**
- Trustworthy AI
- Research and Innovation Action
- 1 October 2022 – 30 September 2025
- Project ID: 101070162

ULTIMATE project: Ambition

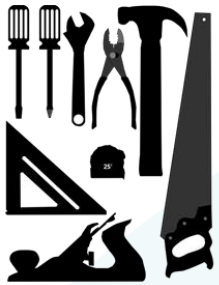
Hybrid AI

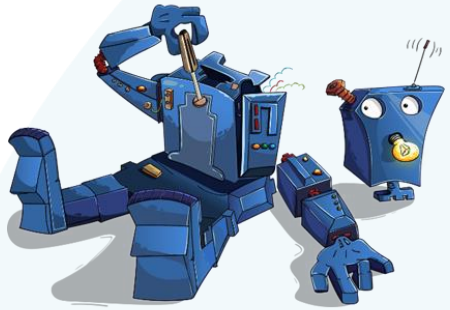


Develop **innovative architectures** to construct and train hybrid AI algorithms

Design & Development

Design **rigorous evaluation methodologies** with appropriate properties





Evaluation

Ensure the **ethical compliance**

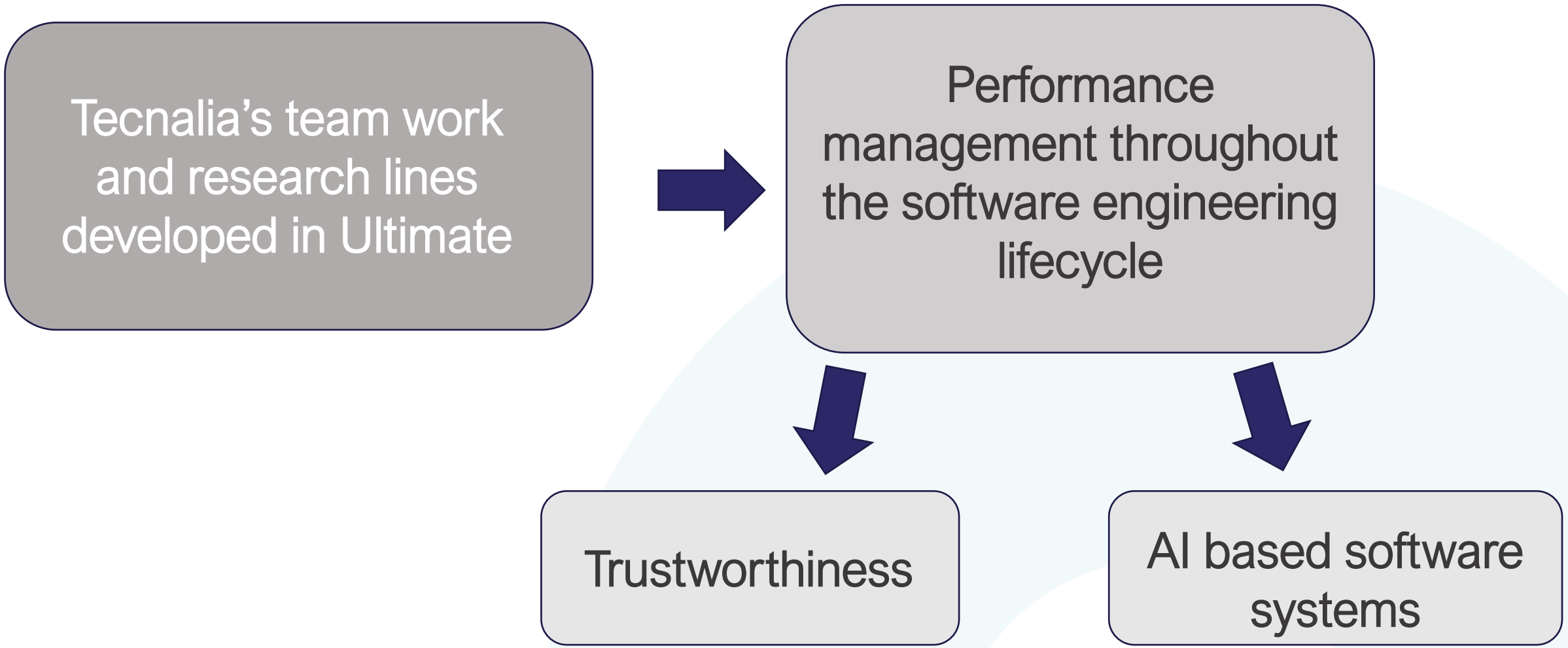
Trustworthiness

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

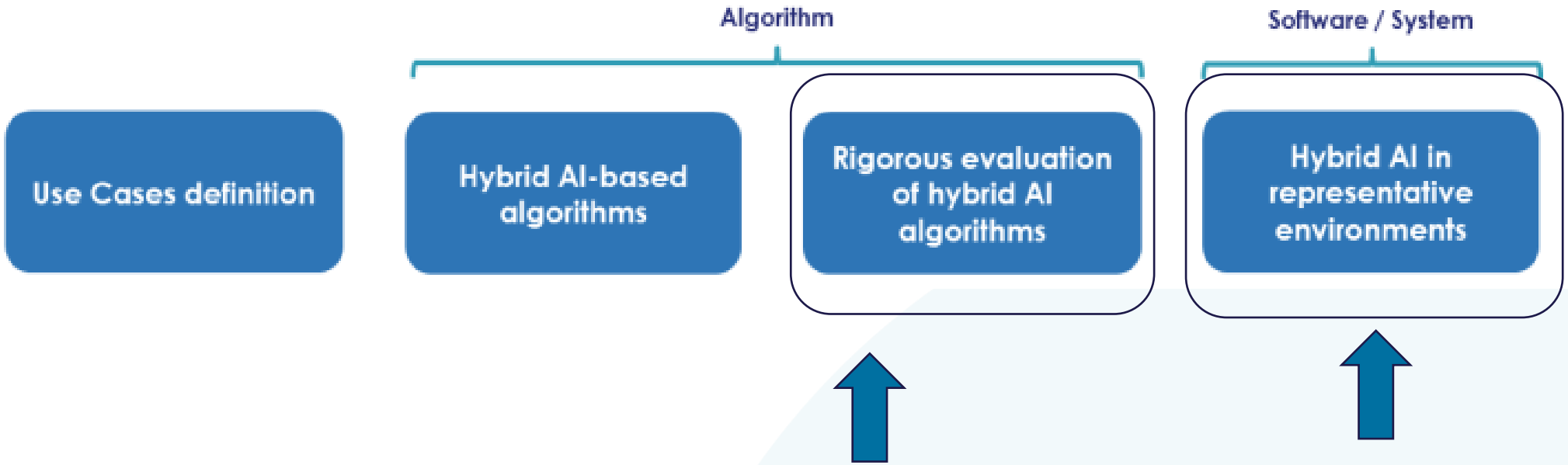
Pre-industrialisation

ULTIMATE project: Consortium

Tecnalia‘s objectives in ULTIMATE (1/2)



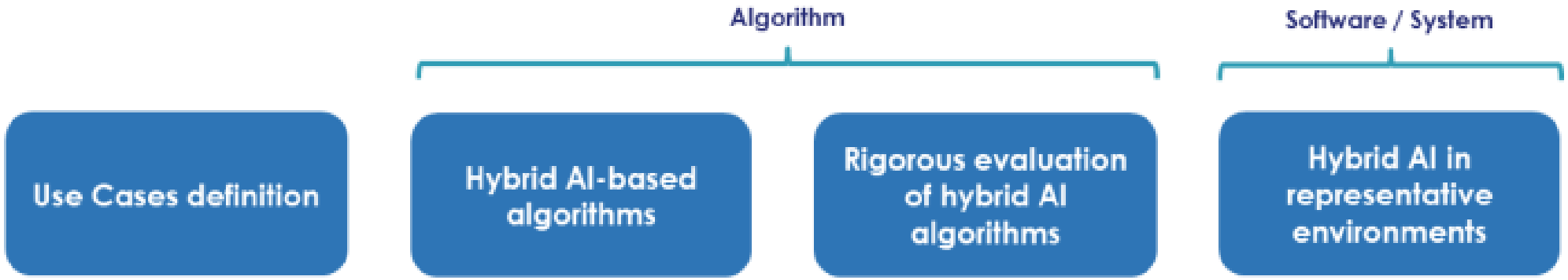
Tecnalia's objectives in ULTIMATE (2/2)



- Identify and use **assessment mechanisms to understand and quantify trustworthiness of hybrid AI models.**
- Identify and select **meaningful performance indicators supporting decision making.**

- Identify and use **assessment mechanisms to understand and quantify AI based system trustworthiness.**
- Identify and select **meaningful performance indicators supporting decision making throughout the engineering lifecycle.**
- **Look for statistical models relating performance indicators** at different system levels and lifecycle points.

Work already started – 1st iteration



TW requirements identification and prioritisation

- Inputs for establishing the scope: AI High Level Expert Group [1], IEEE Ethically Aligned Design [2]
- Approach used: Value Sensitive Design [3]

TW KPIs identification

- Approach being used: Goal Question Metric [4]

LOOKING TO IA ACT [5]

TW = Trustworthiness

AI trustworthy requirements (1/2)

1. Human agency and oversight

- Human agency and autonomy
- Human oversight

2. Technical robustness and safety

- Accuracy / Correctness
- Awareness of misuse
- Controllability
- Reliability
- Reproducibility
- Resilience
- Robustness
- Safety
- Security

3. Privacy and data governance

- Access to data
- Privacy
- Data protection
- Data quality and more

4. Transparency

- Communication
- Explainability / Interpretability / V&V
- Predictability
- Traceability
- Transparency

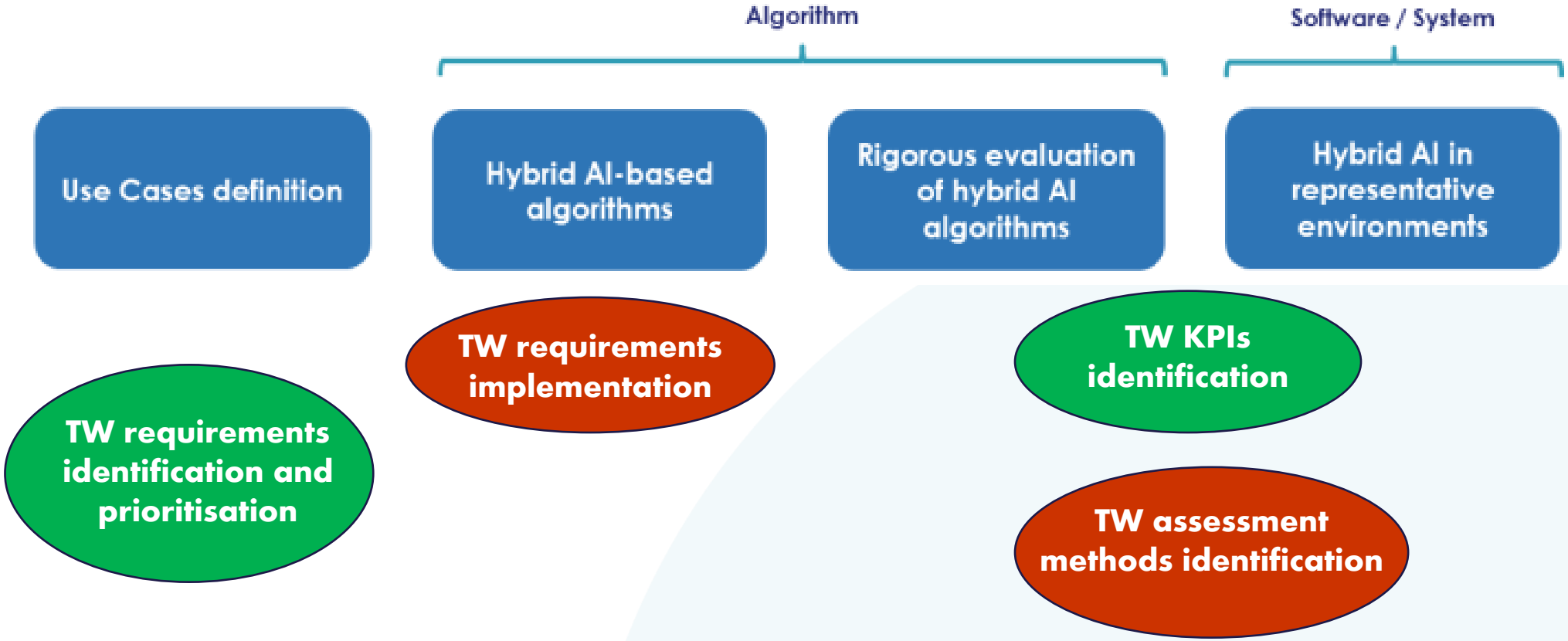
**** Left red border** = Not used on ULTIMATE/VSD*

AI trustworthy requirements (2/2)

5. Diversity, non-discrimination & fairness	6. Societal and environmental wellbeing	7. Accountability
Accessibility and usability	Environmental wellbeing / Sustainability	Accountability
Avoidance of unfair bias	Impact on work and skills / Social impact	Auditability
Fairness	Impact on society / Society and democracy	Risk management
Stakeholder participation		Redressability
		Trade-offs

**** Left red border** = Not used on ULTIMATE/VSD*

Future work



LOOKING TO IA ACT [5]

TW = Trustworthiness

References

[1] EU Independent High-Level Expert Group on Artificial Intelligence (2019) Ethics Guidelines for Trustworthy AI, available at https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
[2] IEEE Ethically Aligned Design v2.0 (for public discussion) https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf
[3] Taebi B., and all (2014) Responsible innovation as an endorsement of public values. The need of interdisciplinary research, in Journal of Responsible Innovation, vol 1, 2014, issue 1, pp 118-124.
[4] The Goal Question Metric Approach. Victor Basili, Gianluigi Caldiera, H. Dieter Rombach. https://www.cs.umd.edu/users/mvz/handouts/gqm.pdf
[5] Artificial Intelligence Act. (21 April 2021). "Proposal for a regulation of the European Parliament and the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts." *EUR-Lex - 52021PC0206* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELLAR:e0649735-a372-11eb-9585-01aa75ed71a1.

Thank you very much for your attention !

mUlti-Level Trustworthiness to IMprove the Adoption of hybrid arTificial intelligence

https://ultimate-project.eu/

nuria.quintano@tecnalia.com

Funded by the European Union

This project has received funding from the European Union's Horizon Europe programme under grant agreement number 1070162.