TAILOR

Foundation of Trustworthy AI:
Integrating Learning, Optimisation and Reasoning

Francesca Pratesi
Institute of Science and Technology, Italian National Council of Reseach
francesca.pratesi@isti.cnr.it

https://tailor-network.eu/

TAILOR is an ICT-48 Network of AI Research Excellence Centers funded by
EU Horizon 2020 research and innovation programme GA No 952215

TAILOR – Unique Selling Point

Actively **bringing together** communities, especially in **reasoning and learning**, in an **academic-industrial** network with the **vision** and **capability** of developing the **scientific foundations** for realising the **European vision** of human-centred **Trustworthy AI**.

This project is funded by the EC under H2020 ICT-48

Ethics Guidelines for Trustworthy AI – Overview

TAILOR
https://tailor-network.eu/

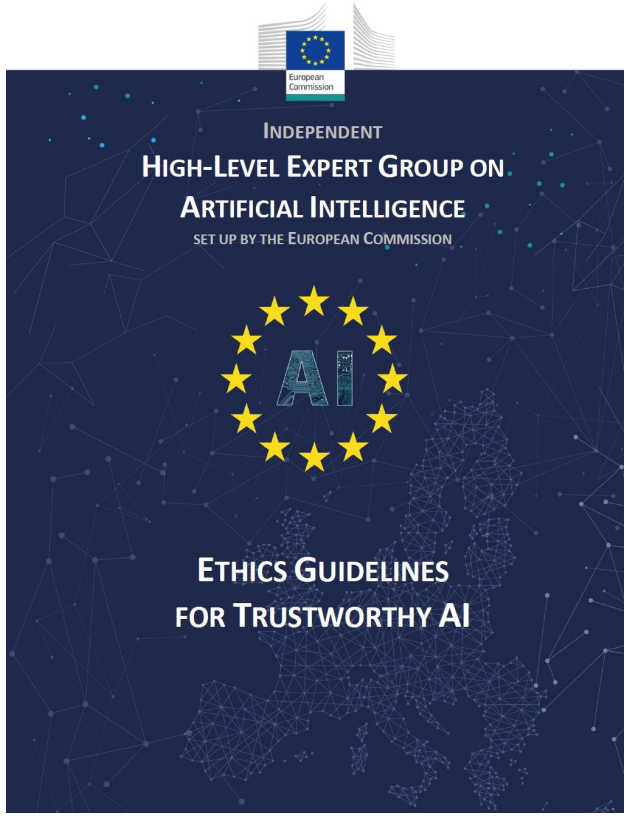
Human-centric approach: AI as a means, not an end

Trustworthy AI as our foundational ambition, with three components

- Lawful AI
- Ethical AI
- Robust AI

Three levels of abstraction

- from principles (Chapter I)
- to requirements (Chapter II)
- to assessment list (Chapter III)



https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

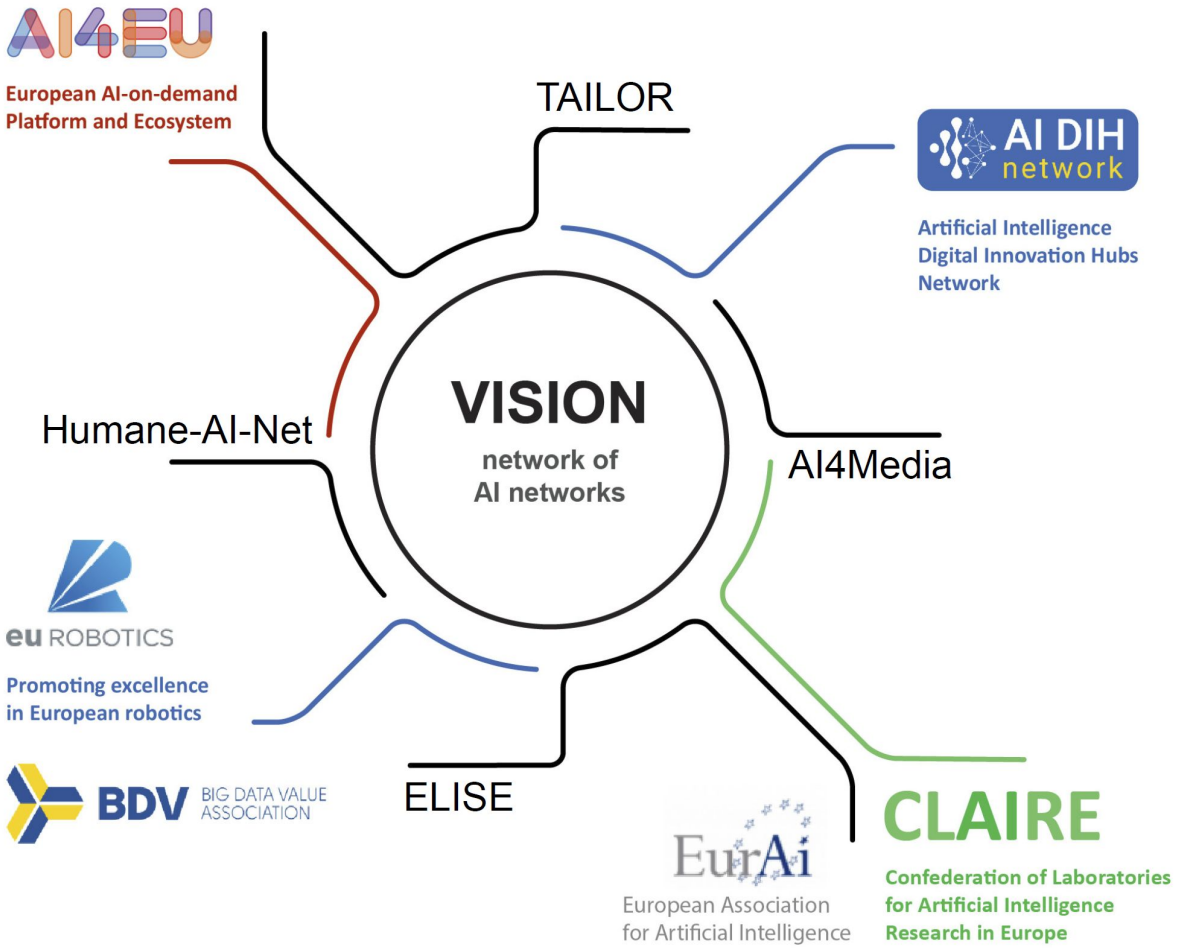
https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

This project is funded by the EC
under H2020 ICT-48

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche

Major EU Projects and Initiatives

- AI4EU
- ICT-48 Networks (4 RIAs + 1 CSA
  - AI4Media
  - ELISE
  - Humane-AI-Net
  - TAILOR
  - VISION (CSA)
- PPP on AI, Data, and Robotics
- Digital Innovation Hubs



This project is funded by the EC under H2020 ICT-48

TAILOR

Foundation of Trustworthy AI:
Integrating Learning, Optimisation and Reasoning

Overview of TAILOR
Francesca Pratesi

TAILOR is an ICT-48 Network of AI Research Excellence Centers funded by
EU Horizon 2020 research and innovation programme GA No 952215

Boosting Capacity to Tackle Major Scientific Challenges

- *How will you boost the research capacity of the AI excellence research centres in Europe by tackling major scientific or technological challenges and by putting in place mechanisms for more efficient collaboration?*
- A **core network** of outstanding AI research centres and major European companies (partners) plus **mechanisms for extending** the network (network members and connectivity fund) to be adaptive and inclusive.
- Five **virtual research environments** to address the **major scientific challenges** required to achieve Trustworthy AI supported by **AI-based** network **collaboration tools**.
- **Strategic** research and innovation **roadmap** to drive the long-term **scientific vision** combined with **bottom-up coordinated actions** collaboratively addressing specific research questions.

This project is funded by the EC
under H2020 ICT-48

Scientific Research and Innovation Roadmap

Identify the promising scientific and technological challenges, from

Curiosity driven (boosting the research capacity)

- A questionnaire to the Extended Roadmap Editorial Board
- Dedicated sessions in the workshops of the scientific WPs

Application driven (engaging industry)

- Outputs of the Theme Development Workshops

This project is funded by the EC
under H2020 ICT-48

TAILOR
https://tailor-network.eu/

TAILOR Consortium

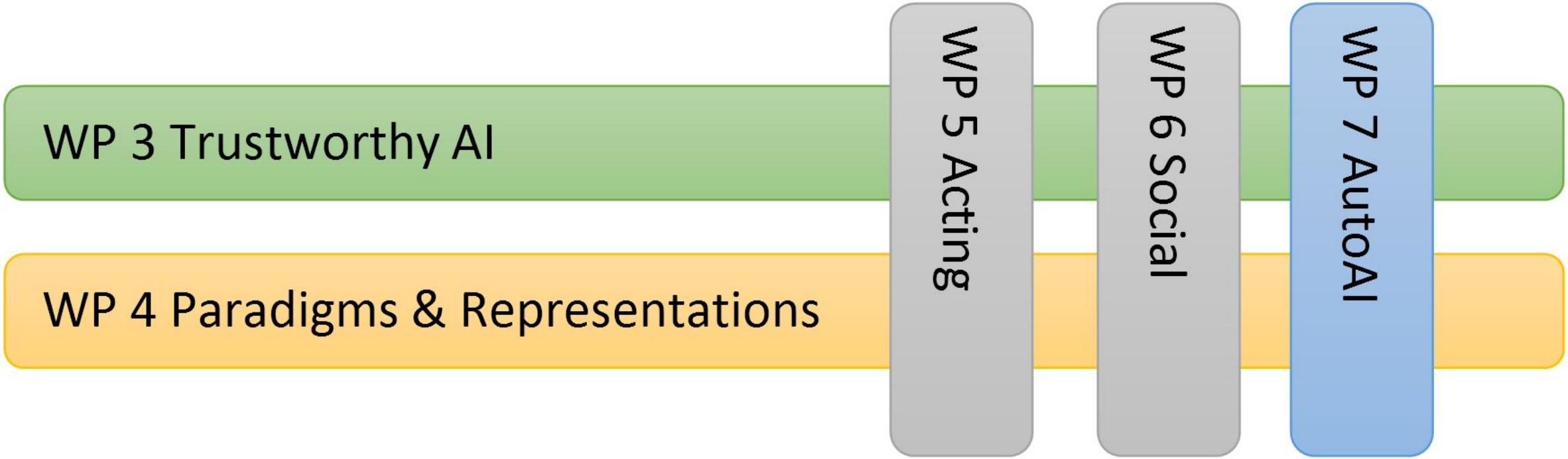
- 54 partners from 18 EU countries (AT, BE x2, CZ x2, DE x8, ES x4, FI, FR x6, GR, IE, IT x8, LU, NL x6, PL, PT, SE x2, SI, SK, UK x4), Israel and Switzerland x2.
- More than 60 network members.
- 23 Core partners (LiU, CNR, INRIA, UCC, KUL, UOR, LEU, IST-UL, UPF, UNIBO, BIU, TUE, CNRS, JSI, TUDA, UNIBRIS, ALU-FR, UOX, UNITN, DFKI, EPFL, FBK, CINI)
- 21 Partners (VUB, CUNI, CEA, CRIL, CVUT, TUD, FhG, TU Graz, IIIA-CSIC, LIRA, UOA, NEO-UMA, PUT, RWTH, slovak.AI, TNO, UniPI, UGA, UNIBAS, UPV, ICL)
- 10 Industry partners (VW, ENG, Tieto, Philips, EDF, ABB, ZF, LIH, CBS, Bosch)

This project is funded by the EC under H2020 ICT-48

CLAIRE

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche

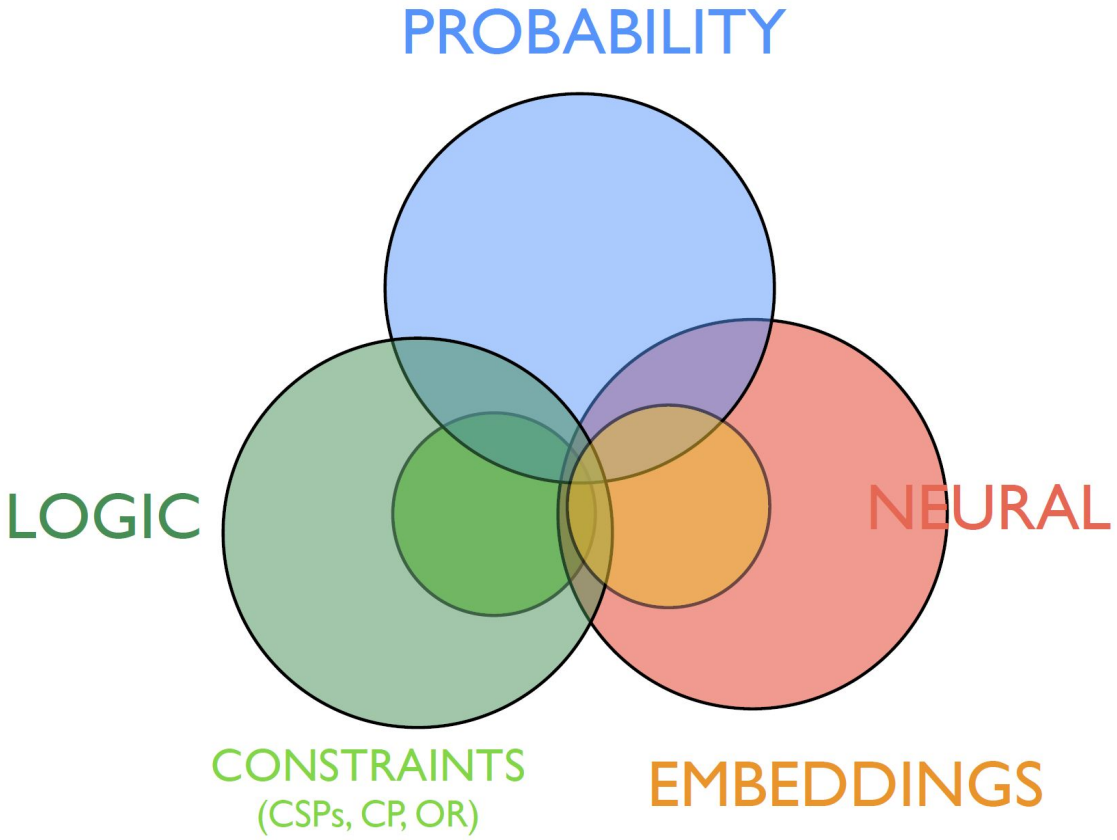
TAILOR Basic Research Program



This project is funded by the EC
under H2020 ICT-48

WP4: Paradigms and Representations

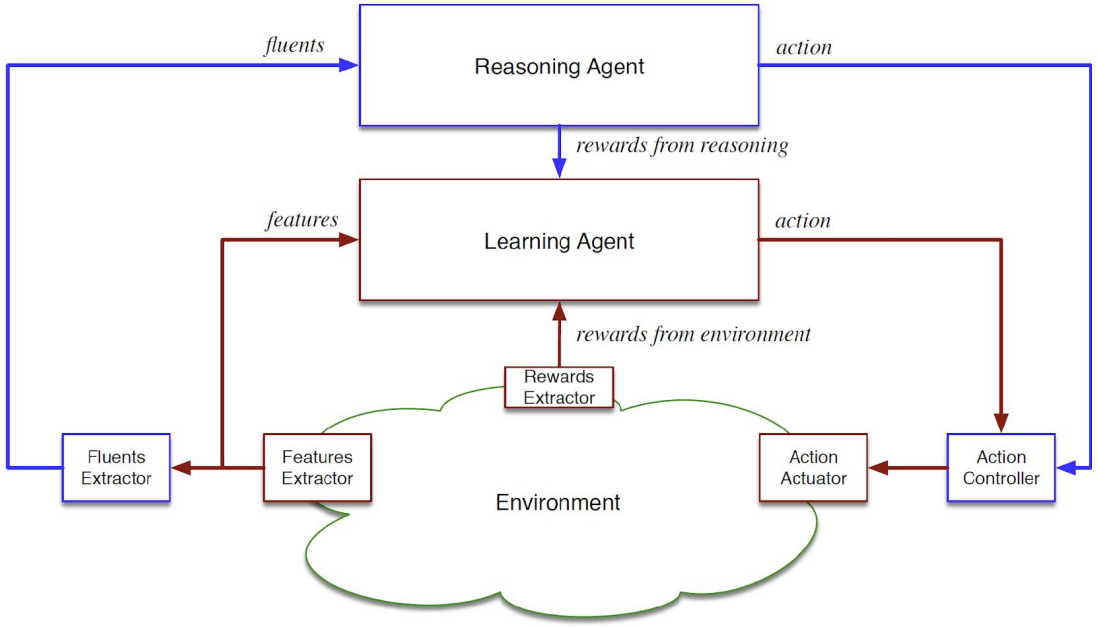
- **Integrated representations** for learning and reasoning
- **Integrated approaches** to learning and optimization
- **Learning and reasoning** with embeddings, knowledge graphs, & ontologies
- Learning and reasoning for perception, spatial reasoning, and vision



This project is funded by the EC
under H2020 ICT-48

WP5: Acting

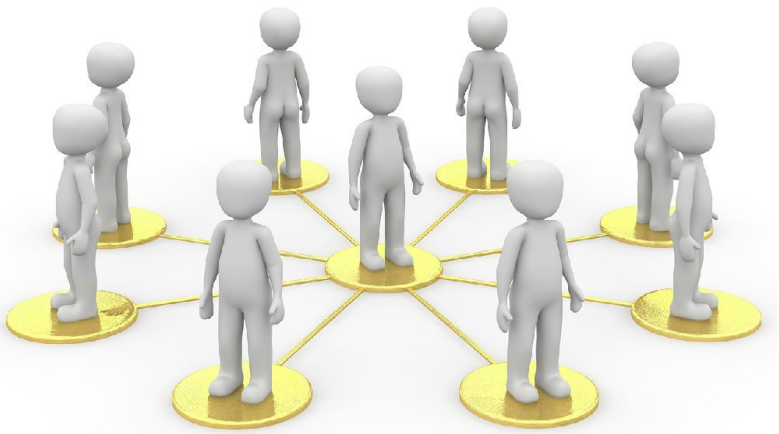
- Extended and multi-facet **models of the world dynamics** and tasks
- Integrating data-based methods with model-based methods in **deciding and learning how to act**
- Learning for reasoners and planners, and reasoners and planners for learning
- Monitoring and controlling to **make actions** AI trustworthy in the real world



This project is funded by the EC under H2020 ICT-48

WP6: Learning and Reasoning in Social Contexts

- Modelling social cognition, collaboration and teamwork
  - **Integrate individual knowledge with knowledge available** to and from other agents
  - Designing social AI systems
- Theoretical models for cooperation between agents
  - **Collaborative decision** making by social agents
  - Aggregate and mediate preferences of multiple agents fairly
  - **Motivate self-interested agents** to execute their tasks and **towards the greater good**
- Learning from others
  - **Social learning**
- Emergent behaviour, agent societies and social networks
  - Designing **complex social structures**, organizations and institutions



This project is funded by the EC
under H2020 ICT-48

WP7: AutoAI

- **Automate labour-intensive, error-prone aspects of building AI systems** to make them more trustworthy and robust
  - AutoML in the wild
  - Beyond standard supervised learning
  - Self-monitoring AI Systems
  - Multi-objective AutoAI
  - Ever-learning AutoAI

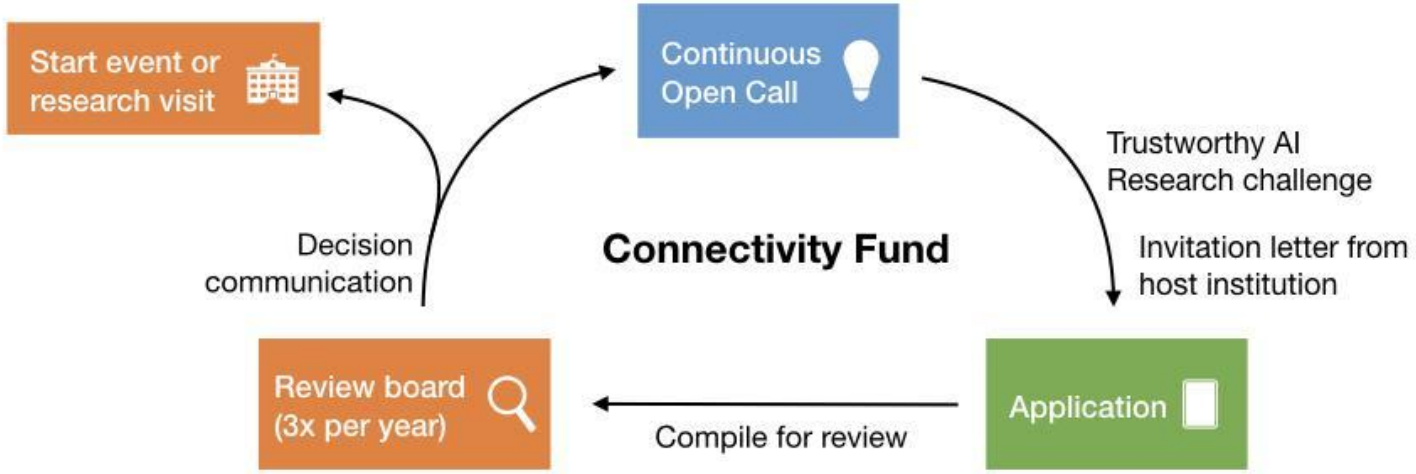
This project is funded by the EC
under H2020 ICT-48

TAILOR
https://tailor-network.eu/

Connectivity Fund

Call closes Nov 15!

- 1.5 million EUR fund, third-party funding (guest or host is non-TAILOR)
- Open call, reviewed every 4 months (March, July, November)
  - Submitted by non-TAILOR host or guest
  - Max. 60.000 EUR per visit/workshop, covers travel, housing, and sustenance



Research Visits

We support research visits between 1 and 12 months. We will pick up the bills so that you can focus on doing excellent AI. You must either be from a non-TAILOR lab visiting a TAILOR lab, or vice versa.

Workshops

We support workshops that bring people all across Europe together to solve hard problems in an open atmosphere. Workshops should explicitly bring TAILOR and Non-TAILOR researchers together.

This project is funded by the EC under H2020 ICT-48

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" Consiglio Nazionale delle Ricerche

ACM Computing Survey - Special Issue

ACM Journals

CALL FOR PAPERS

ACM Computing Surveys
Special Issue on Trustworthy AI

Guest Editors:

- **ROBERTA CALEGARI**, Alma Mater Studiorum–Università di Bologna, Italy, roberta.calegari@unibo.it
- **FOSCA GIANNOTTI**, Scuola Normale Superiore, Italy, fosca.giannotti@isti.cnr.it
- **MICHELA MILANO**, Alma Mater Studiorum–Università di Bologna, Italy, michela.milano@unibo.it
- **FRANCESCA PRATESI**, National Research Council, Italy, francesca.pratesi@isti.cnr.it

Trustworthy Artificial Intelligence (TAI) systems have become a priority for the European Union and have increased their importance worldwide. The European Commission has consulted a High-Level Expert Group that has delivered a document on Ethics Guidelines for Trustworthy AI to promote Trustworthy AI principles.

TAI has three overarching components, which should be met throughout the system's entire life cycle: (1) it should be lawful, complying with all applicable laws and regulations, (2) it should be ethical, ensuring adherence to ethical principles and values, and (3) it should be robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm. Each component in itself is necessary but not sufficient for the achievement of TAI. Ideally, all three components work in harmony and overlap in their operation. If, in practice, tensions arise between these components, society should endeavor to align them.

From a practical perspective these principles boil down into TAI dimensions, including:

- robustness
- reproducibility and replicability
- safety
- transparency and explainability
- diversity, non-discrimination and fairness
- safety
- privacy and data governance
- sustainability
- accountability

- **Important Dates**
- Open for Submissions: July 31, 2022
- Submissions deadline: September 15, 2022
- First-round review decisions: December 20, 2022
- Deadline for revision submissions: February 26, 2023
- Notification of final decisions: April 16, 2023
- Tentative publication: Fall 2023

This project is funded by the EC under H2020 ICT-48

TAILOR

Foundation of Trustworthy AI:
Integrating Learning, Optimisation and Reasoning

The dimensions of Trustworthy AI

Francesca Pratesi

TAILOR is an ICT-48 Network of AI Research Excellence Centers funded by
EU Horizon 2020 research and innovation programme GA No 952215

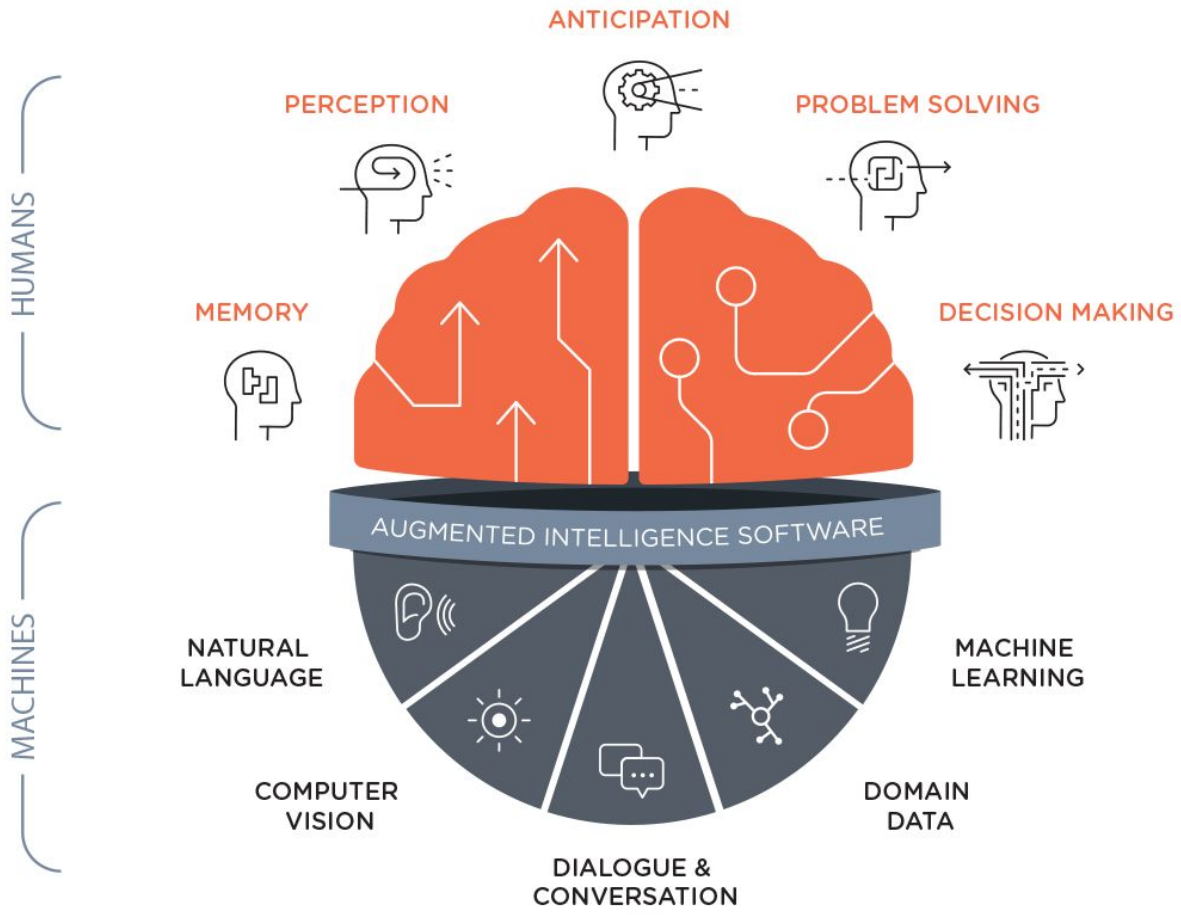
Trustworthy AI

- Goal
  - establish a continuous interdisciplinary dialogue for investigating methods and methodologies
  - ***"To create AI systems that incorporate trustworthiness by design"***
- Organized along the 6 dimensions of Trustworthy AI:
  - Explainability,
  - Safety and Robustness,
  - Fairness,
  - Accountability,
  - Privacy, and
  - Sustainability
- One transversal task that links the 6 dimensions among and ensures coherence and coordination across the activities.

This project is funded by the EC under H2020 ICT-48

Why is Artificial Intelligence Different

- Scale
- Speed
- Single-mindedness
- Optimization-based
- Cannot break the rules
- No needs
- No real consequences or “skin in the game”



This project is funded by the EC under H2020 ICT-48

What (ethical) decisions?

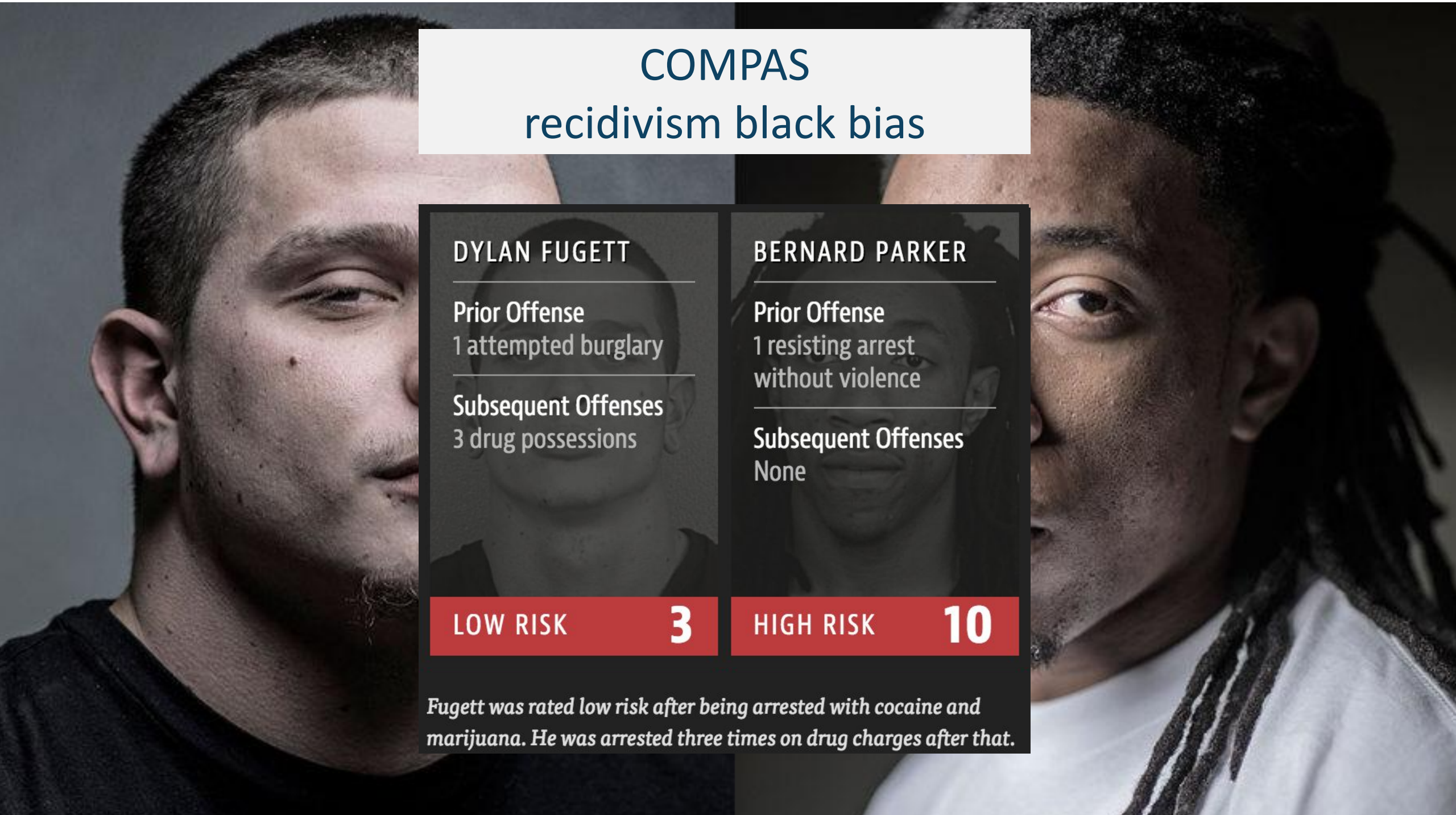
- Can we teach ethics to AI?
- **Should** we teach ethics to AI?
- Understanding ethics
  - Which values? Whose values?
  - Who gets a say?
- Using ethics
  - What is proper action given values?
  - Are ethical theories of use?
  - How to prioritise values?
- Being ethical
  - Is knowing ethics enough?

This project is funded by the EC under H2020 ICT-48

WP3: Explainable AI Systems

This project is funded by the EC under H2020 ICT-48

COMPAS
recidivism black bias



What is a Black Box Model?

A ***black box*** is a model, whose internals are either unknown to the observer or they are known but uninterpretable by humans.

Example:

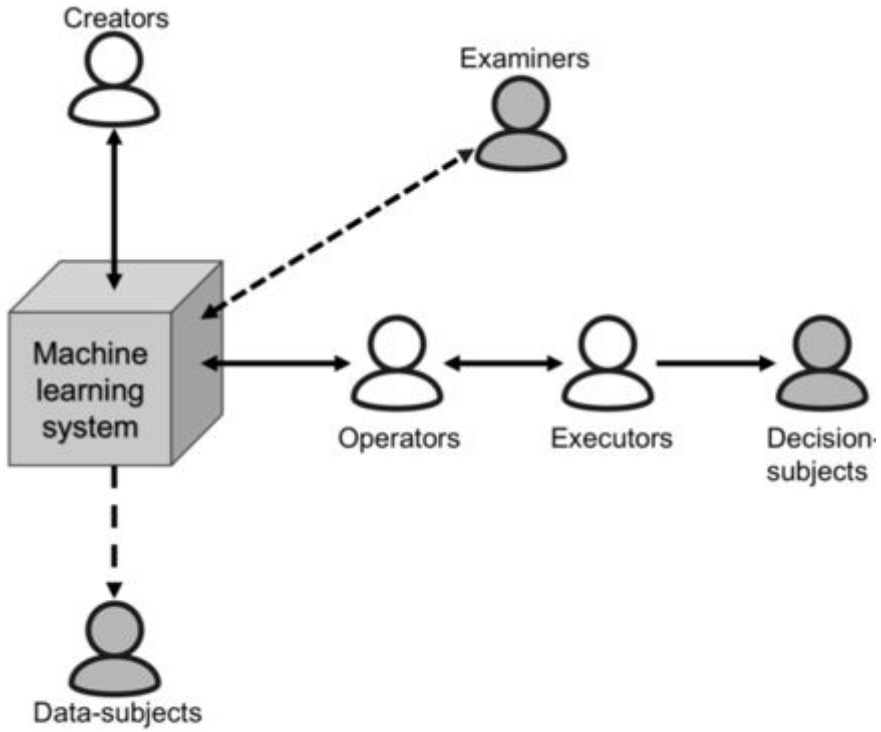
- DNN
- SVM
- Ensemble

Guidotti, Monreale, Ruggieri, Turini, Giannotti, Pedreschi. *A survey of methods for explaining black box models.* ACM Computing Surveys (CSUR), *51*(5), 93 (2018).

This project is funded by the EC
under H2020 ICT-48

What is a good explanation, and for whom

- **End users** "Am I being treated fairly?"
  "Can I contest the decision?"
  "What could I do differently to get a positive outcome?"
- **Engineers, data scientists**: "Is my system working as designed?"
- **Regulators** " Is it compliant?"

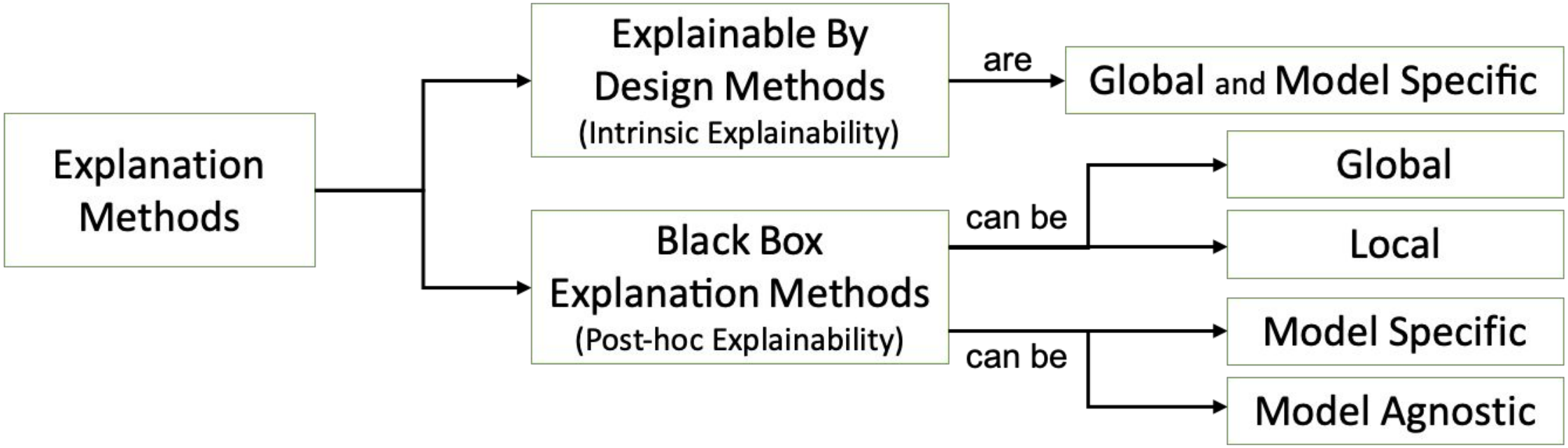


An ideal explainer should model the *user background*

[Tomsett et al. 2018, Weld and Bansal 2018, Poursabzi-Sangdeh 2018, Mittelstadt et al. 2019]

This project is funded by the EC under H2020 ICT-48

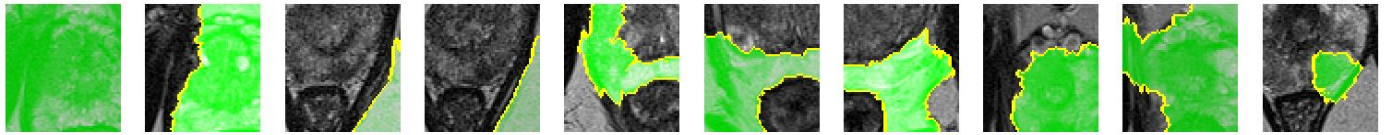
A way to think about Explainability



This project is funded by the EC
under H2020 ICT-48

Research challenges on Explainable AI

- We would like to apply, to the **same case study**, different Explainability techniques (e.g., LIME, LORE, ABELE, CRUSADE), providing differents outputs, such as:
  - saliency maps (possibly with integrated gradients to be more precise):



  - factual and counterfactual rules

r = { 29 > 0.50, 46 > 0.50, 47 > 0.50, 25 > 0.50, 58 > 0.50, 56 > 0.50, 40 > 0.50 } --> { class: 1 }
c = { { 46 <= 0.50 } --> { class: 0 }, { 40 <= 0.50 } --> { class: 0 }, { 58 <= 0.50 } --> { class: 0 } }

  - exemplars and counter-exemplars images
- Then, we must to find a reliable way to **compare the results**

Guidotti, Monreale, Ruggieri, Pedreschi, Turini, Giannotti. *Local Rule-Based Explanations of Black Box Decision Systems*. arXiv preprint arXiv:1805.10820(2018)
Guidotti, Monreale, Giannotti, Pedreschi, Ruggieri, Turini. *Factual and Counterfactual Explanations for Black Box Decision Making*. IEEE Intelligent Systems, 34(6)
Guidotti, Monreale, Matwin, Pedreschi, *Black Box Explanation by Learning Image Exemplars in the Latent Feature Space*. ECML PKDD 2019

This project is funded by the EC
under H2020 ICT-48

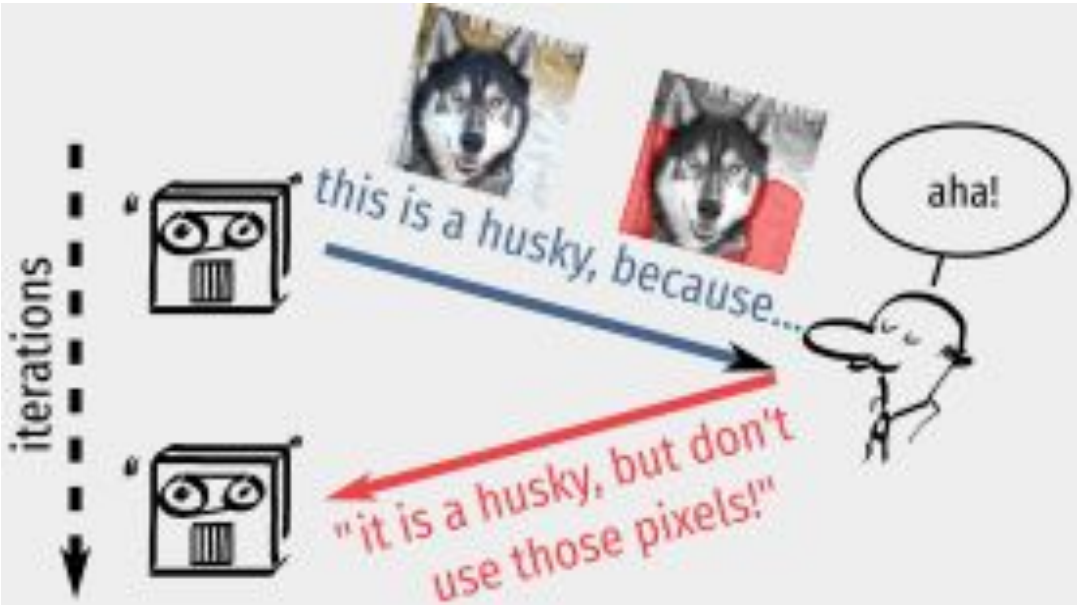
Research challenges on Explainable AI

- Apply Explainability techniques to new domains
  - E.g., Cybersecurity Threat Detection, where there exist a number of effective classification algorithms
  - The starting points: analyze processes (accessed resources, used libraries) to discover promptly malwares and security threats
  - testing different algorithms able to provide explanations,
  - reviewing other “classic” methods and their possible extensions

This project is funded by the EC
under H2020 ICT-48

Research challenges on Explainable AI

- Explanatory Active Learning/Human-Machine Interaction
  - How to combine interaction and explanations to provide better results?



https://github.com/stefanoteso/awesome-explanatory-supervision
(Also workshop on Interactive Machine Learning @ AAAI-22)

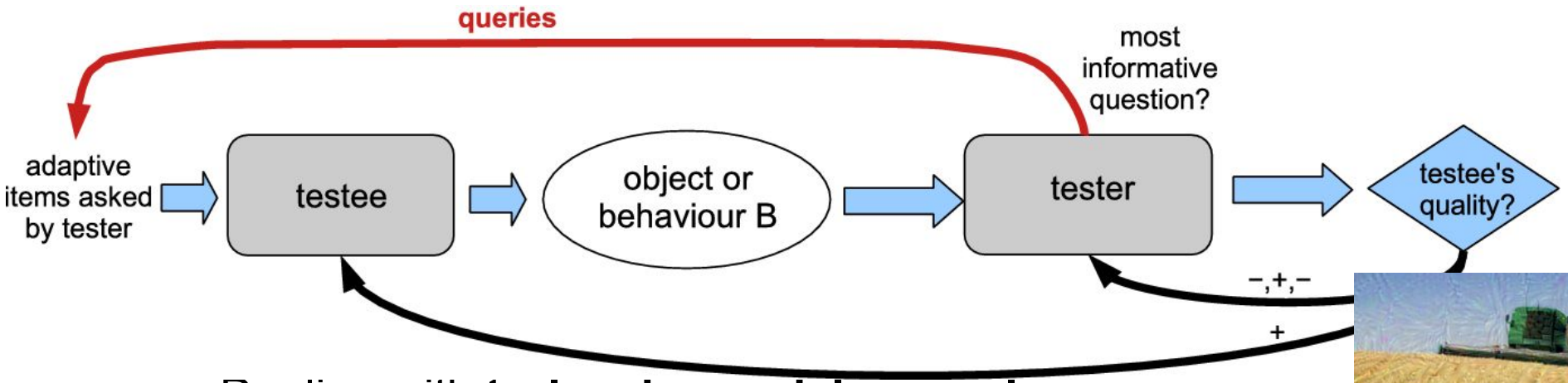
This project is funded by the EC
under H2020 ICT-48

WP3: Safety and Robustness

This project is funded by the EC under H2020 ICT-48

Research challenges on Security

- **Robust Evaluation**: prevent specialisation and test replacement
  - Adversarial testing / Use of adaptive testing

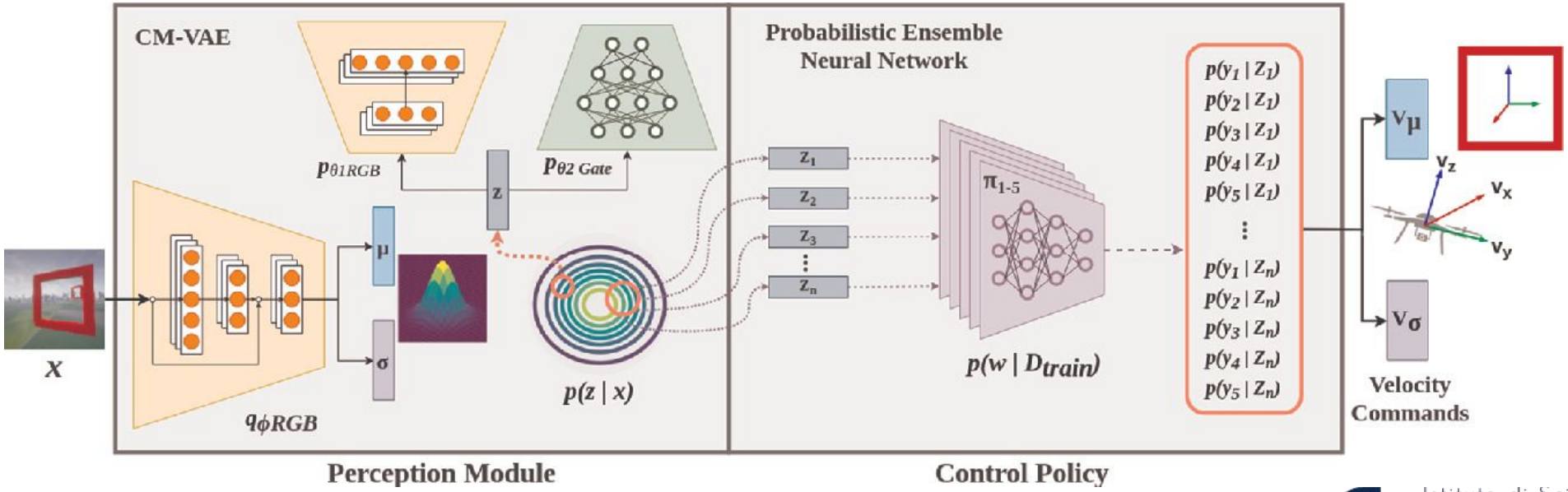


- Dealing with **truly adversarial examples**:
  - Extend adversarial scenarios adding the difficulty condition using, e.g., item response theory.
  - Study the detection of (non-)adversarial perturbed examples by understanding their discrimination characteristics.

This project is funded by the EC under H2020 ICT-48

Research challenges on Security

- From safety critical systems to **artificial general intelligence (AGI) risks**
  - AGI is the hypothetical ability of an AI system to understand or learn any intellectual task that a human being can.
- **Dynamic Risk Assessment** of AI
  - ensuring system is capable of managing risks in relevant operational situations
  - with DNN, some residual risk must be managed at runtime



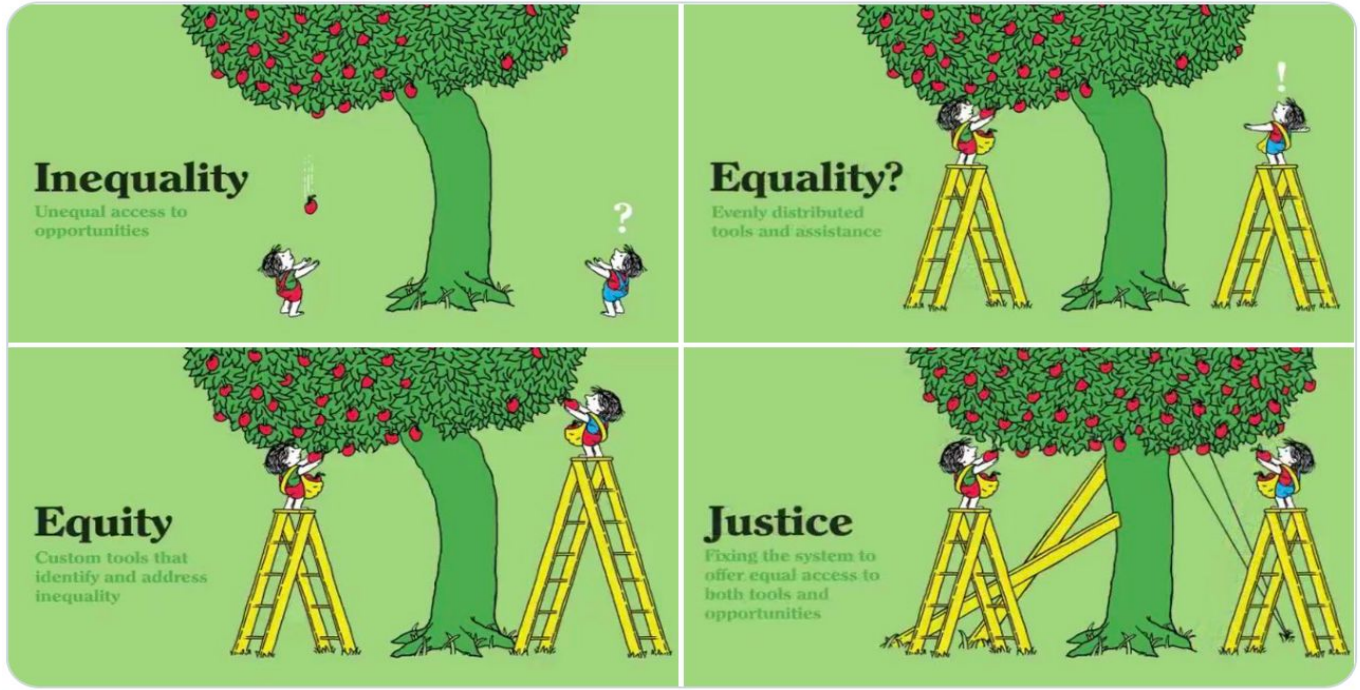
This project is funded by the EC under H2020 ICT-48

WP3: Fairness, Equity, and Justice by design

This project is funded by the EC under H2020 ICT-48

What is Fairness and Fairness-Aware ML?

- Impartial and just treatment or behavior without favoritism or discrimination.



- *Fairness-aware machine learning* algorithms seek to provide methods under which the predicted outcome of a classifier operating on data about people is fair or non-discriminatory

This project is funded by the EC
under H2020 ICT-48

Research challenges on Fairness, Equity, and Justice

- Operationalizing Fairness Metrics. The case of credit scoring
  - From a legal and ethical point, fairness often requires non-discrimination, which requires some form of equal treatment, or lack of bias
  - But more generally, fairness is about how people deserve to be treated, and sometimes equal treatment is unfair
    - E.g., positive discrimination is sometimes warranted or even required, e.g. to deal with historical injustices or vulnerable groups
  - Fairness metrics often focus only on one aspect or dimension, different fairness metrics may conflict, and the metrics focus on what *can* be measured.
  - Operationalizations may not fully cover the intended fairness notion or concern

This project is funded by the EC
under H2020 ICT-48

WP3: Accountability and Reproducibility by design

This project is funded by the EC under H2020 ICT-48

Accountability

Accountability regards the decisions on the attribution of **moral responsibility** and **who is responsible** (and **in what sense**) for untoward outcomes.

AI systems offer unprecedented opportunities to develop faster and more reliable systems, also helping humans to take decisions, but they also introduce more gray areas.

European Group on Ethics in Science and New Technologies, *Artificial Intelligence, Robotics and 'Autonomous' Systems*, Brussels 9 March 2018, doi:10.2777/531856

This project is funded by the EC under H2020 ICT-48

Reproducibility

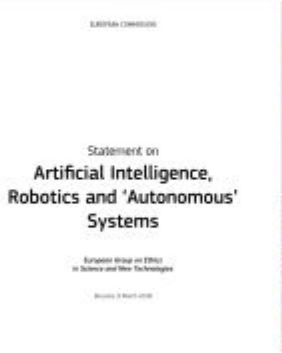
- **Reproducibility of methods**: the ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results
- **Reproducibility of results**: the production of corroborating results in a new study, having used the same experimental methods
- **Reproducibility of inference**: the drawing of qualitatively similar conclusions from either an independent replication of a study or a reanalysis of the original study

This project is funded by the EC
under H2020 ICT-48

Research challenges on Accountability and Reproducibility

TAILOR
https://tailor-network.eu/

- Propose methods to **model and analyze** emergent effects and diffusion of responsibility in **multi-agent settings** (problem of many hands and many things)



- Consider simultaneously multiple dimensions which have notable **value tensions** (e.g., privacy and security), especially in opaque and non-reproducible applications.
- Take a **multidisciplinary perspective** on evaluating how to avoid accountability gaps given a broad spectrum of dimensions (e.g., explainability, robustness, and fairness).
- **Evaluate metadata models** on a list of competency questions identified as pivotal for ensuring reproducibility and accountability. It will identify gaps in the schemas and how these schemas might complement each other to document the different dimensions of trustworthiness

This project is funded by the EC
under H2020 ICT-48

WP3: Respect for Privacy

This project is funded by the EC under H2020 ICT-48

Respect for Privacy

This project is funded by the EC under H2020 ICT-48

Research Challenge on Privacy

- There is still the need to investigate new methodologies and approaches for:
  - automatically detecting **privacy risks**
  - designing robust **data anonymization algorithms**
  - designing AI algorithms that respect **by design** privacy constraints
  - collaborative projects between academia and **industry** for a continuous transfer of anonymization techniques

PbD
www.privacybydesign.ca

This project is funded by the EC
under H2020 ICT-48

Research Challenge on Privacy

- Study the tensions and the **interplay with other aspects and human values**

This project is funded by the EC
under H2020 ICT-48

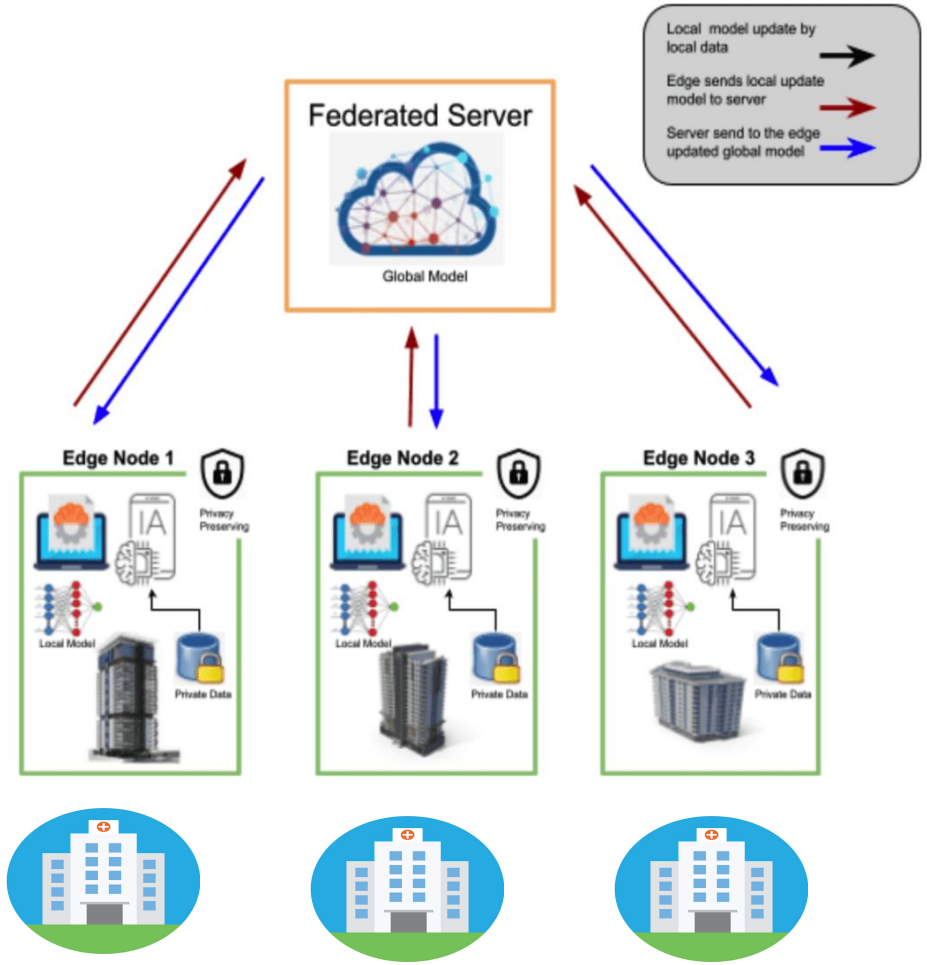
Research Challenge on Privacy

Provide global or local **explanations on privacy risks**

This project is funded by the EC under H2020 ICT-48

Research Challenge: Federated Learning

"Federated Learning is a ML setting where multiple distributed parties, called clients, under the orchestration of a main server, cooperate to train a shared global model, while keeping their data private"



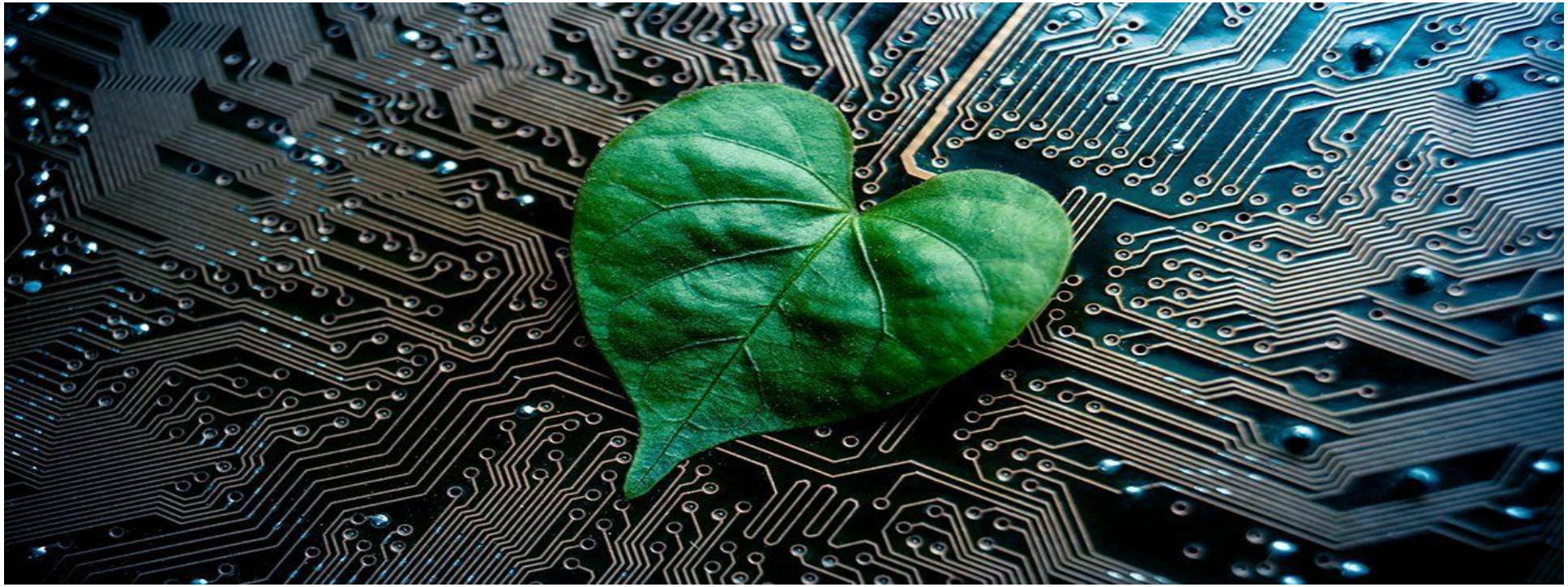
This project is funded by the EC under H2020 ICT-48

WP3: Sustainability

This project is funded by the EC under H2020 ICT-48

Sustainability

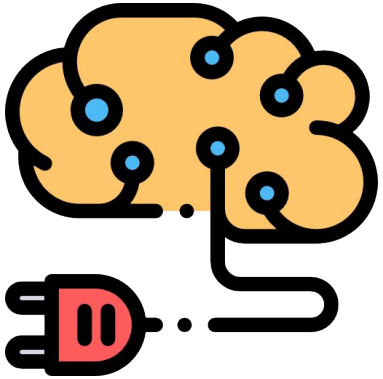
Energy consumption and carbon footprint of AI systems and solution like deep networks is extremely high.



This project is funded by the EC
under H2020 ICT-48

Research Challenge on Sustainability

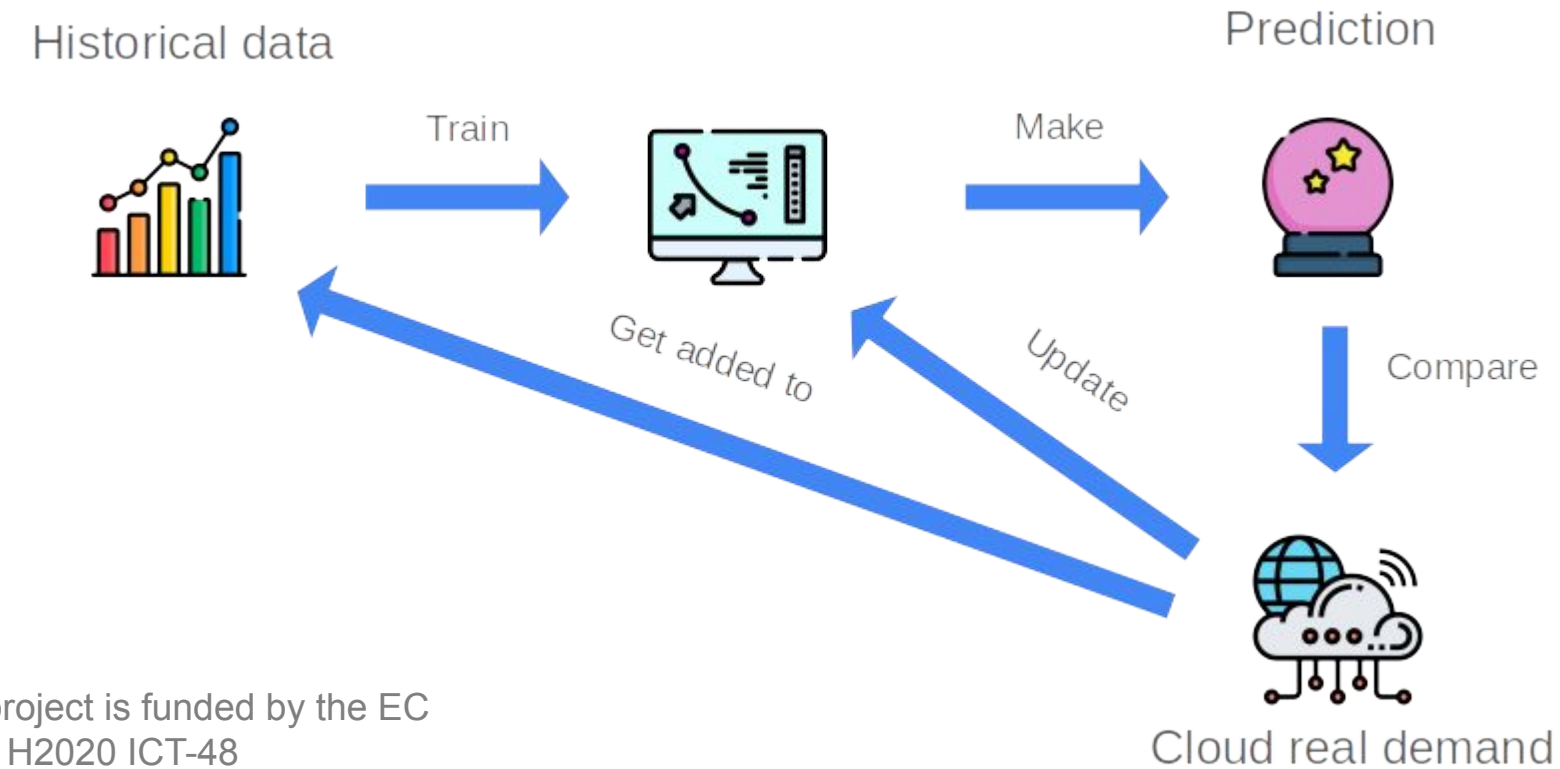
- Reduce the energy needed to train AI solutions
  - Making them learning faster
  - Recycle learning from different applications (transfer learning)
- Optimise the data centres in which the AIs are trained
  - Reduce the waste of resources



This project is funded by the EC
under H2020 ICT-48

Research Challenge on Sustainability

- Predict the expected behavior of a deep NN both at training and inference time, to steer the **dynamic reconfiguration and adaptation** of the training/inference process



This project is funded by the EC
under H2020 ICT-48


thank you