Trustworthy Human-Centric AI

The European Approach

Fredrik Heintz
Dept. of Computer Science, Linköping University
fredrik.heintz@liu.se
@FredrikHeintz

LINKÖPINGS UNIVERSITET

A European approach to Artificial Intelligence

European Commission

Ethics Guidelines for Trustworthy AI – Overview

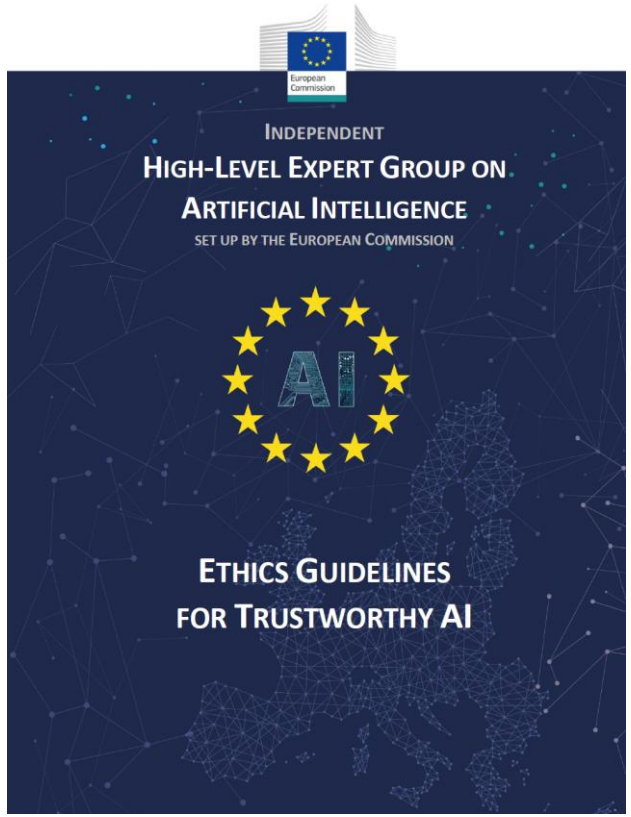
Human-centric approach: AI as a means, not an end

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

Ethics Guidelines for Trustworthy AI – Principles

4 Ethical Principles based on fundamental rights

Respect for human autonomy

Augment, complement and empower humans

Prevention of harm

Safe and secure. Protect physical and mental integrity.

Fairness

Equal and just distribution of benefits and costs.

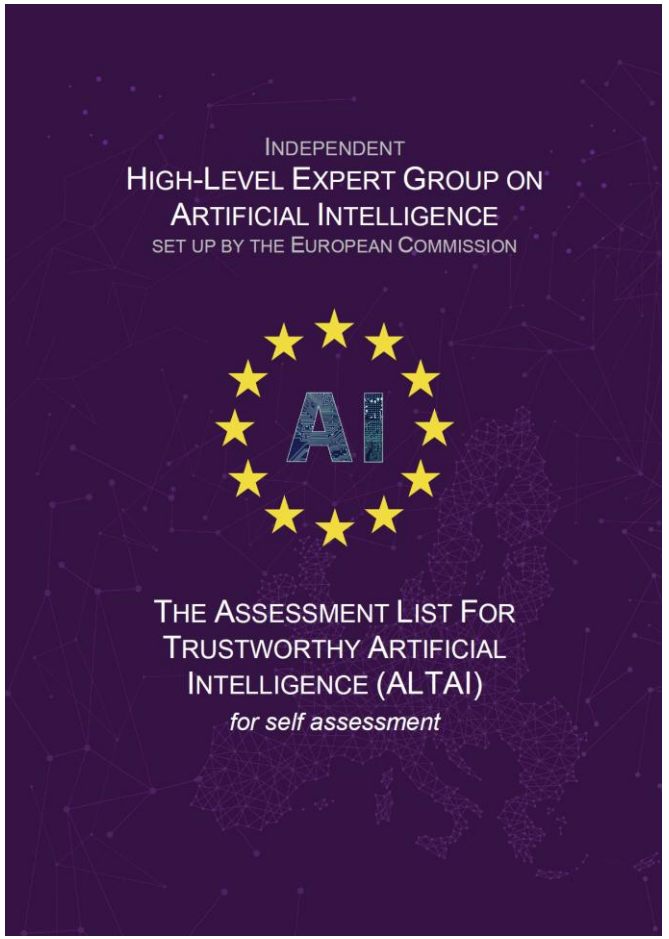
Explicability

Transparent, open with capabilities and purposes, explanations

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

The Assessment List for Trustworthy AI (ALTAI)

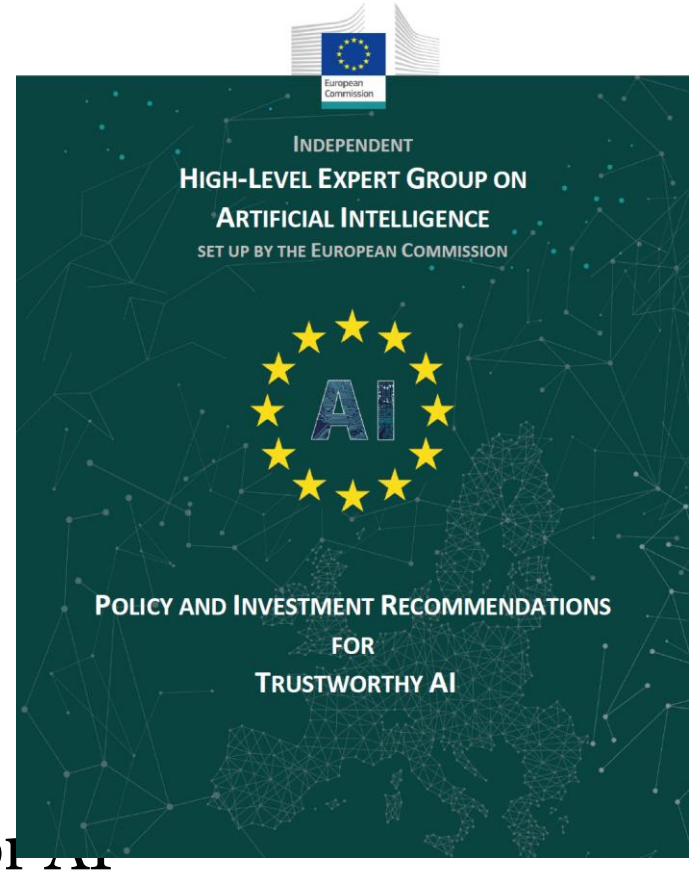
- **REQUIREMENT #1 Human Agency and Oversight**
  - Human Agency and Autonomy; Human Oversight
- **REQUIREMENT #2 Technical Robustness and Safety**
  - Resilience to Attack and Security; General Safety; Accuracy; Reliability, Fall-back plans and Reproducibility
- **REQUIREMENT #3 Privacy and Data Governance**
  - Privacy; Data Governance
- **REQUIREMENT #4 Transparency**
  - Traceability; Explainability; Communication
- **REQUIREMENT #5 Diversity, Non-discrimination and Fairness**
  - Avoidance of Unfair Bias; Accessibility and Universal Design; Stakeholder Participation
- **REQUIREMENT #6 Societal and Environmental Well-being**
  - Environmental Well-being; Impact on Work and Skills; Impact on Society at large or Democracy
- **REQUIREMENT #7 Accountability**
  - Auditability; Risk Management



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

Policy & Investment Recommendations

- Using AI to build a positive impact in Europe
  - Empowering and Protecting Human and Society
  - Transforming Europe's Private Sector
  - Catalyzing Europe's Public Sector
  - Ensuring World-Class Research Capabilities
- Leveraging Europe's enablers of AI
  - Raising Funding and Investments for AI
  - Building Data and Infrastructure for AI
  - Generating appropriate Skills and Education for AI
  - Establishing an appropriate governance framework for AI



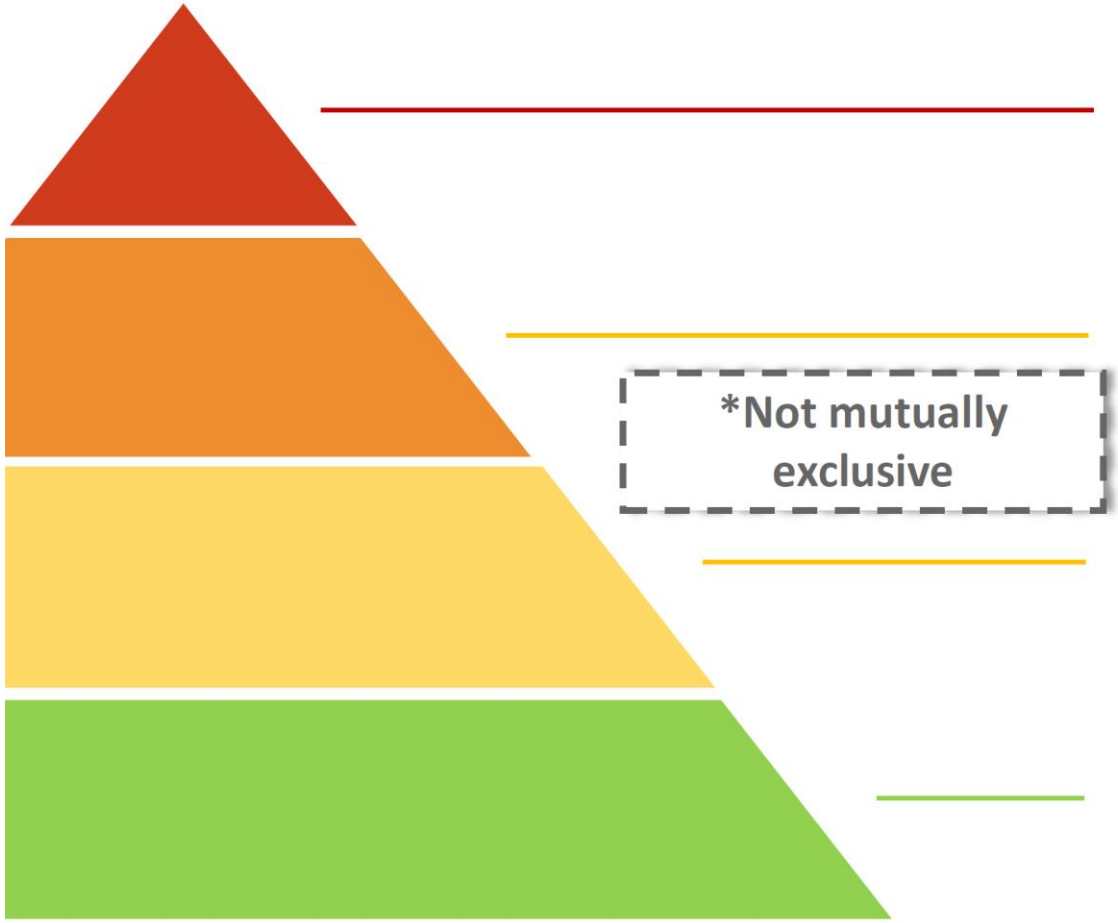
https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence

Sectorial Considerations on the Policy and Investment Recommendations

- **Sectors**
  - The Manufacturing and Industrial IoT Sector
  - Public Sector: the e-Government domain
  - Public Sector: Justice and law-enforcement
  - The Healthcare Sector
- **General comments**
  - The AI HLEG Policy and Investment Recommendations for Trustworthy AI are perceived as important and relevant
  - There is merit in refining the AI HLEG Policy and Investment Recommendations for Trustworthy AI to account for sectoral specificities
  - Trustworthiness is seen as a crucial feature of European AI
  - There is widespread concern about the need to close the skills gap
  - Europe should be a leader in responsible research and innovation in the field of AI
  - Good governance and the widespread sharing of best practices can promote regulatory certainty
  - Data quality, availability and interoperability must be at the core of EU policy

https://futurium.ec.europa.eu/en/european-ai-alliance/blog/sectorial-considerations-trustworthy-ai-taking-ais-context-specificity-account

A risk-based approach



*Not mutually exclusive

Unacceptable risk
e.g. social scoring
— **Prohibited**

High risk
e.g. recruitment, medical devices
— **Permitted** subject to compliance with AI requirements and ex-ante conformity assessment

'Transparency' risk
'Impersonation' (bots)
— **Permitted** but subject to information/transparency obligations

Minimal or no risk
— **Permitted** with no restrictions

European Commission

Requirements for high-risk AI systems (Title III, Chapter 2)

Establish and implement **risk management system**

&

in light of the **intended purpose** of the AI system

- Use high-quality **training, validation and testing data** (relevant, representative etc.)
- Draw up **technical documentation** & set up **logging capabilities** (traceability & auditability)
- Ensure appropriate degree of **transparency** and provide users with **information** on capabilities and limitations of the system & how to use it
- Ensure **human oversight** (measures built into the system and/or to be implemented by users)
- Ensure **robustness**, **accuracy** and **cybersecurity**

Foster an Ecosystem of excellence

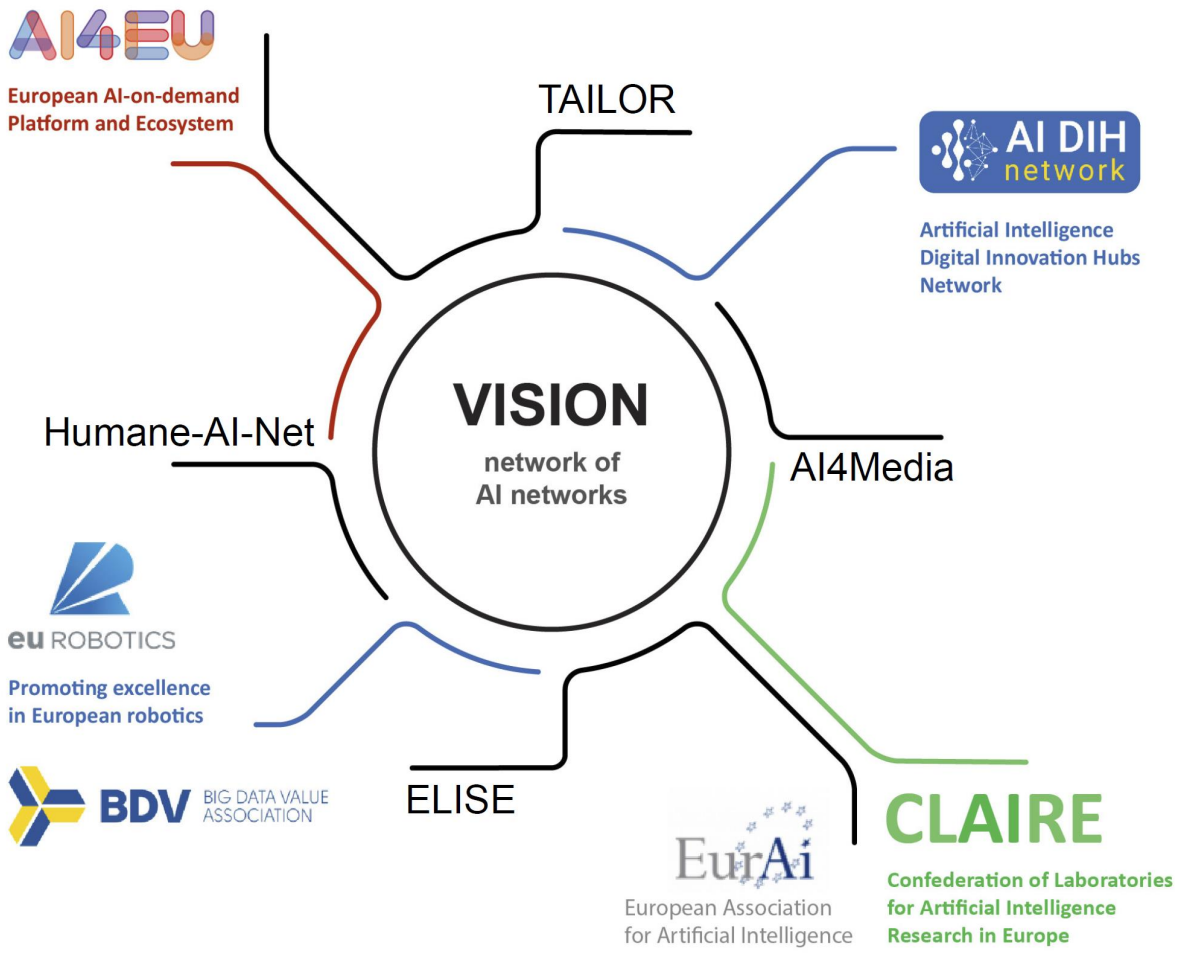
Leverage EU strengths to accelerate AI development, application and use

- Coordinate with Member States
- Support excellence and testing centers
- Networks of leading universities
- Digital Innovation Hubs with high degree of AI expertise
- Scale up equity fund for AI in InvestEU
- New Public Private Partnership in AI
- Adopt AI Programme for public sector

The ecosystem of excellence builds on the Member State AI group's work on the Coordinated Plan on AI (2018).

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

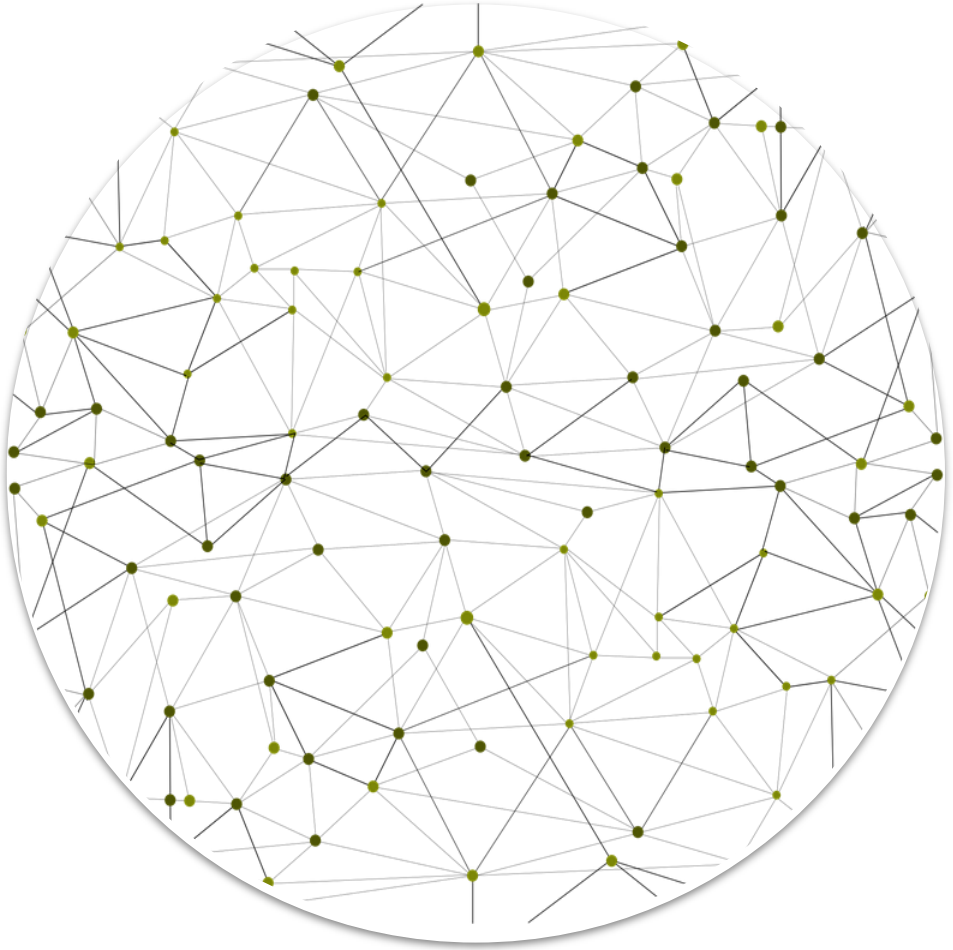



CLAIRE
Confederation of Laboratories for
Artificial Intelligence Research in Europe
Excellence across all of AI. For all of Europe. With a human-centered focus.
www.claire-ai.org

CLAIRE Vision: AI for Good, AI for All

CLAIRE VISION: Research Network

World's largest
AI research network



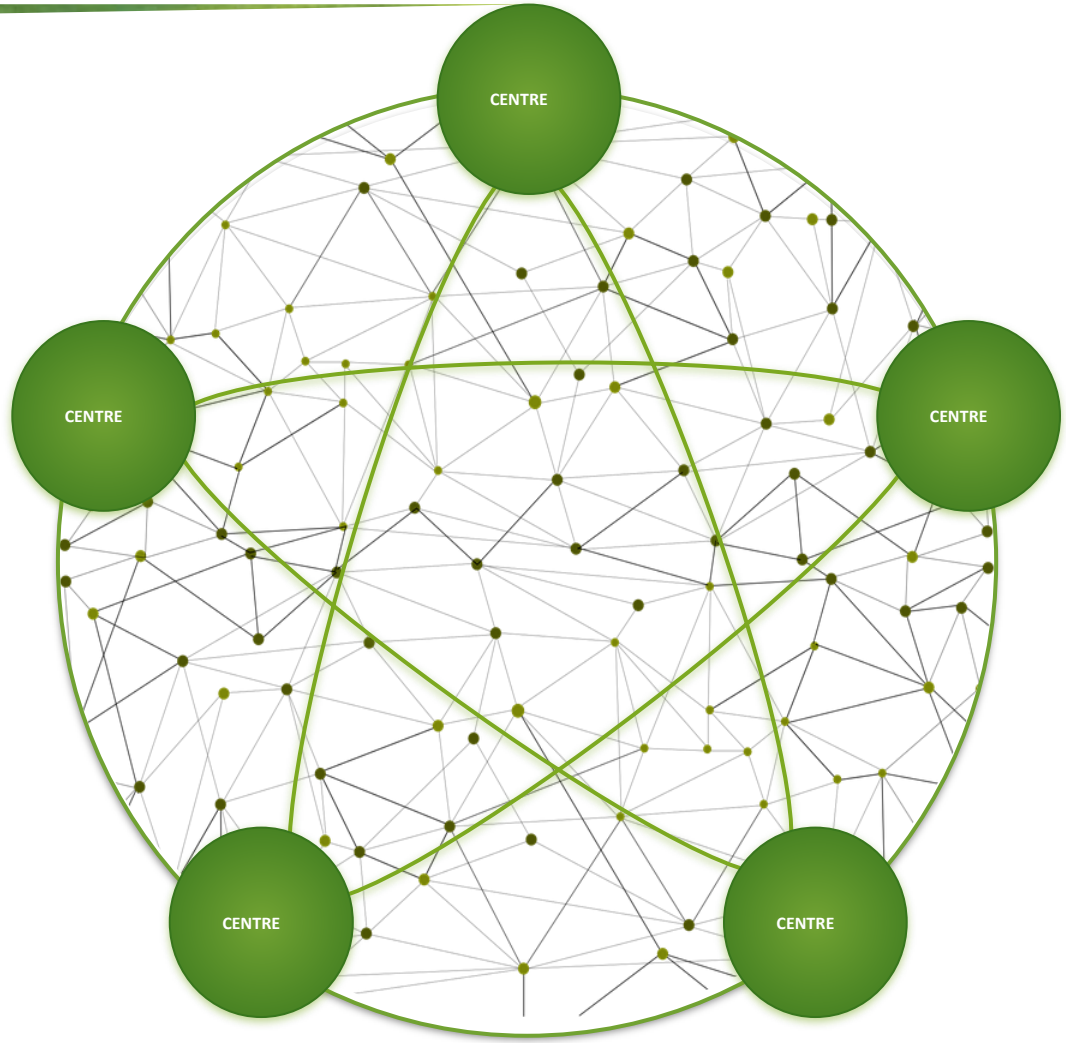
400+ research groups,
labs & organisations
across all of Europe

22 000+ AI researchers
& support staff
in 35 countries

Major stake in
ICT-48 Networks of
Centres of Excellences

CLAIRE VISION: Centres of Excellence

Top researchers, staff & infrastructure

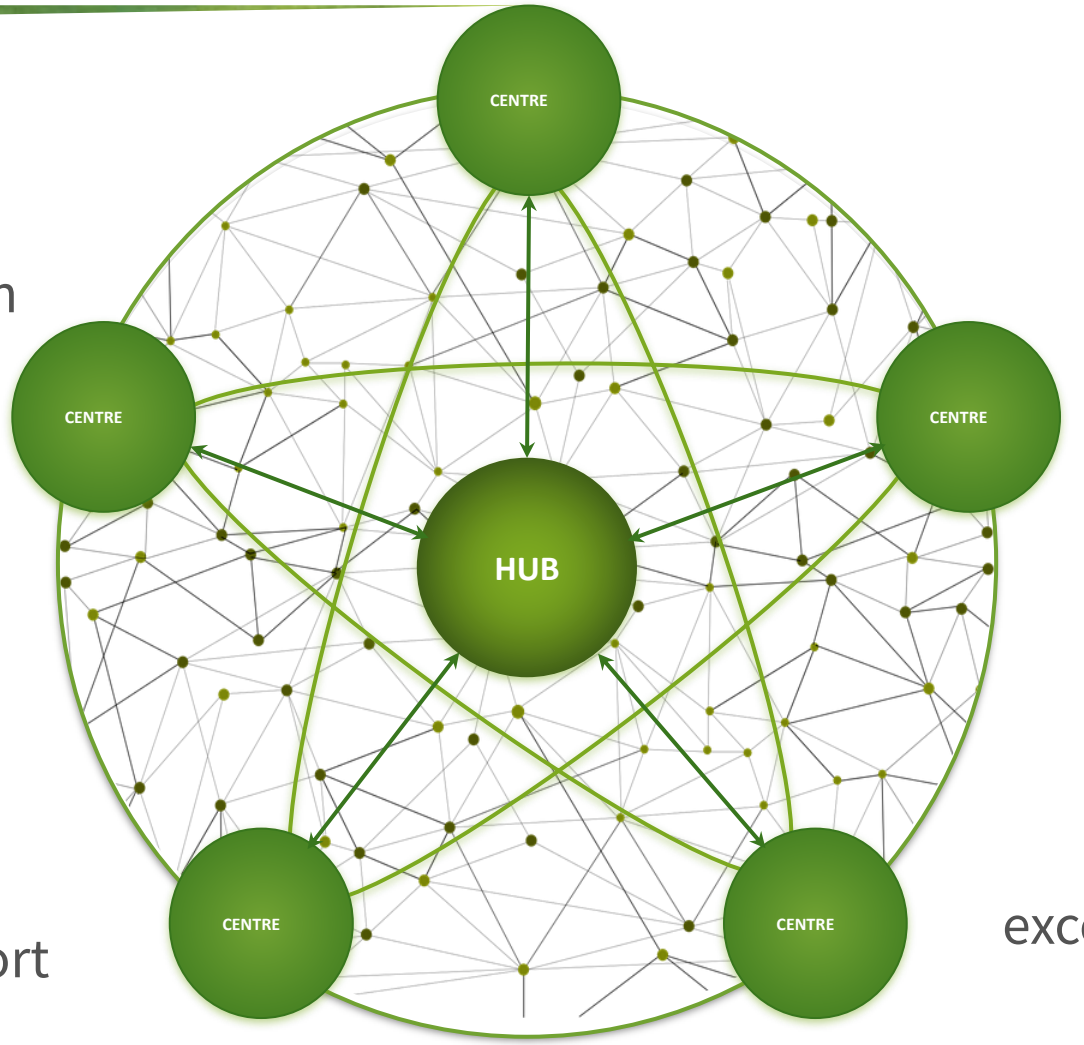


Focused on specific aspects of AI

Regional focal points of AI activity

CLAIRE VISION: Lighthouse Centre ("CERN for AI")

Focal point for exchange & interaction



Global attractor for AI talent

World-leading infrastructure & support

Symbol for European excellence & ambition in AI

Our supporters

- 3737 individual supporters
  - 2161 AI experts
  - 1028 supporters in industry
- International institutions
- Top AI research centre
- 9 EU governments

CLAIRE

- Non-profit organisation (AISBL)
- Goal: positioning Europe as global leader in human-centred Artificial Intelligence (AI)
- CLAIRE Research Network:
  400+ AI research groups & institutes
  representing over 22 000 employees
- CLAIRE Innovation Network:
  leverage this for benefit of Europe's businesses

Our Mission

We are at a crossroads where

1. **Machine learning is at the heart of a technological and societal artificial intelligence revolution** involving multiple sister disciplines, with large implications for the future competitiveness of Europe.
2. **Europe is not keeping up:** many of the top labs, as well as many of the top places to do a PhD, are located in North America; moreover, AI investments in China and North America are significantly larger than in Europe.
3. **the distinction between academic research and industrial labs is vanishing**, with a significant part of the basic research now being done in industry (with substantial research freedom, and higher salaries), rapid commercialization of results, and academic institutions worldwide struggling to retain their best scientists (with negative implications not only for research but also for the education of future

AI DATA ROBOTICS PARTNERSHIP EU

ABOUT PARTNERS SRIDA RESOURCES EVENTS CONTACT

https://ai-data-robotics-partnership.eu/

European Partnership on Artificial Intelligence, Data and Robotics

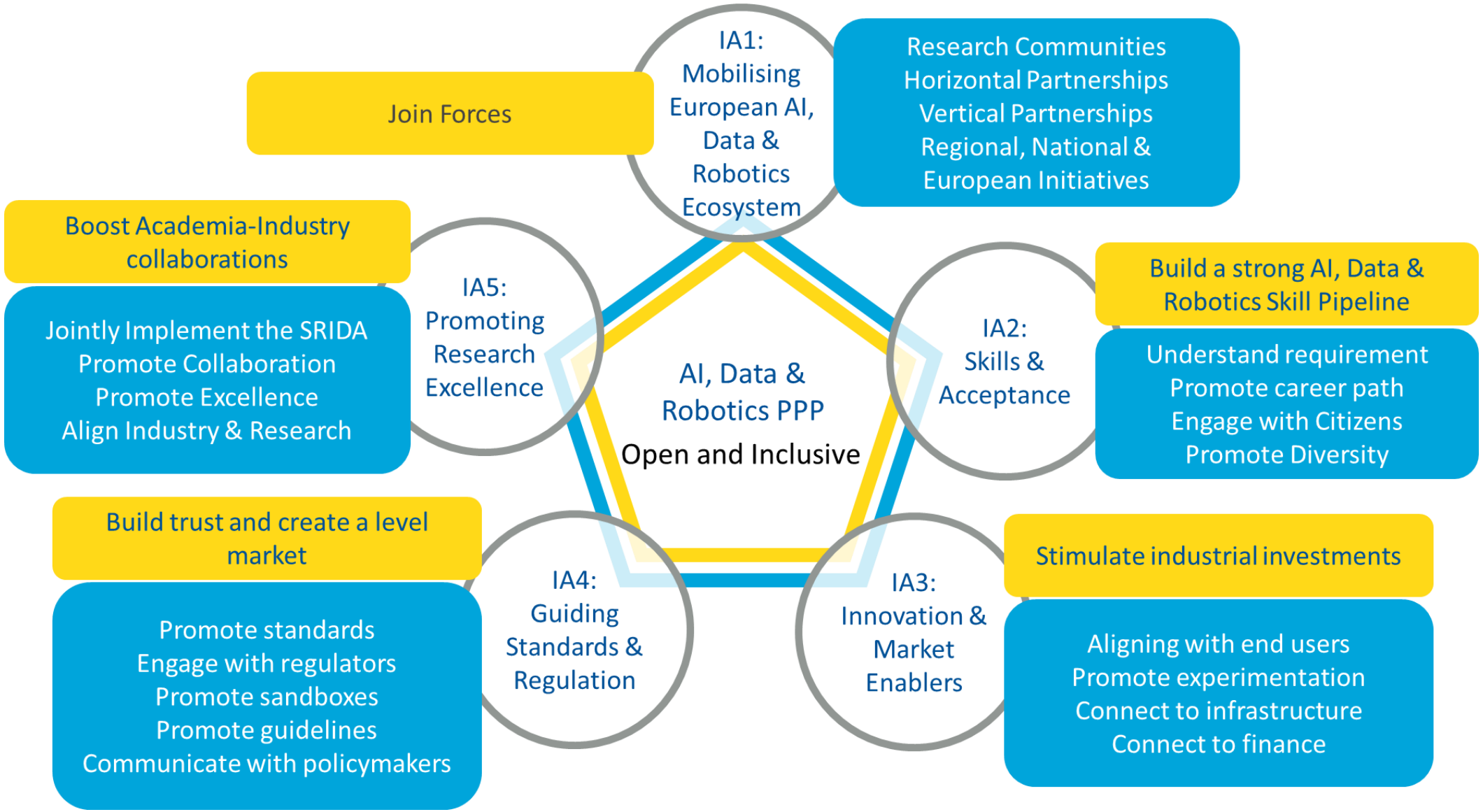
The Vision of the Partnership is to boost European competitiveness, societal wellbeing and environmental aspects to lead the world in researching, developing and deploying value-driven trustworthy AI, Data and Robotics based on fundamental European rights, principles and values.

- **Openness and inclusiveness** to bring European and cross-domain knowledge together to fulfill the vision
- **Openness to include new partners** (new businesses, new experts, new knowledge, new entrepreneurs etc.)
- Joint strategy leveraging **European strengths and unique selling points** to be developed
  BUT also a strong focus on new emerging businesses (e.g. verticals, service businesses etc.)
- Leading principle must **produce valuable content in/for Europe** and to overcome particular interests. **Trust to each others.**
- **Well balanced** amount of members covering AI, Data and Robotics and well balanced between Research and Industry.
  Representativeness of the new innovation forces/communities (e.g Start-ups, high-tech companies, …)

A joint initiative by:

BDV BIG DATA VALUE ASSOCIATION CLAIRE ellis EurAi eu ROBOTICS European Commission

Implementing the Partnerships: key IAs



TAILOR

Foundation of Trustworthy AI:
Integrating Learning, Optimisation and Reasoning

Fredrik Heintz

Dept. of Computer Science, Linköping University

fredrik.heintz@liu.se, @FredrikHeintz

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

LINKÖPINGS UNIVERSITET

Vision

Develop the scientific foundations for Trustworthy AI

integrating learning, optimisation and reasoning.

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

TAILOR ICT-48 Network

TAILOR brings together **leading AI research centres** from **learning**, **optimisation** and **reasoning** together with **major European companies** representing **important industry sectors** into a **single scientific network** addressing the **scientific foundations** of **Trustworthy AI** to **reduce** the **fragmentation**, **boost** the **collaboration**, and **increase** the **AI research capacity** of Europe as well as **attracting and retaining talents** in Europe.

Basic Facts

- Type of action: RIA (Research and Innovation Actions)
- Proposal number: 952215
- Starting date: September 1st 2020
- Duration: 36 months
- # Partners: 54
- Coordinator: Fredrik Heintz, Linköping University (Sweden)
- Total Budget: 12 M€

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

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)

CLAIRE

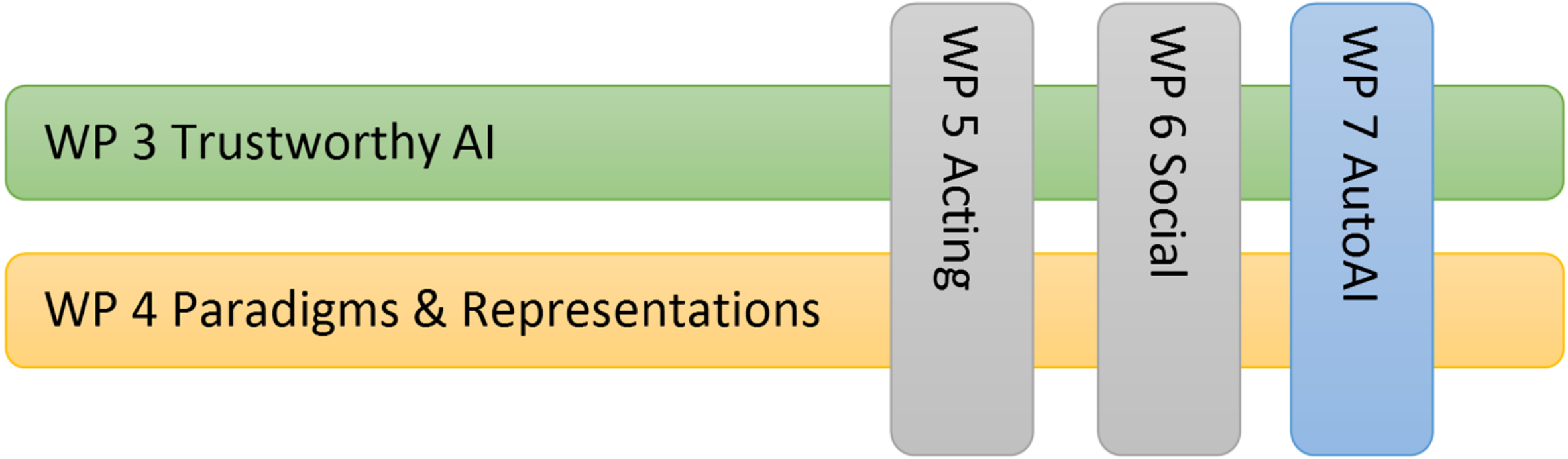
This project is funded by the EC under H2020 ICT-48

Scientific Vision

- **AI should be trustworthy** and developed in a **human-centric** way with the goal of improving individual and societal well-being.
- To be trustworthy AI systems should be **lawful**, **ethical** and **robust**.
- We intend to do the research necessary to **develop the scientific** and **technical foundations** to achieve **trustworthy AI**.
- The ability to **learn**, to **reason** and to **optimize** are central and essential for AI in general and trustworthy AI in particular.
- The network will work to **bridge the gap between learning, reasoning and optimization**, and to **unite** these approaches in **common frameworks** that pave the way towards more powerful trustworthy AI systems.

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

Basic Research Program



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

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

WP3: Trustworthy AI

- **Task 3.1**: Explainable AI Systems
  - Generate multimodal explanations
- **Task 3.2**: Safety and Robustness
  - Make AI systems safe and robust
- **Task 3.3**: Fairness, Equity, and Justice by design
  - Make AI systems fair
- **Task 3.4**: Accountability and Reproducibility by design
  - Acc.: blameworthiness, liability, prevent misuse
  - Rep.: measures, quality standards, and procedures to model the development of learning methods for AI

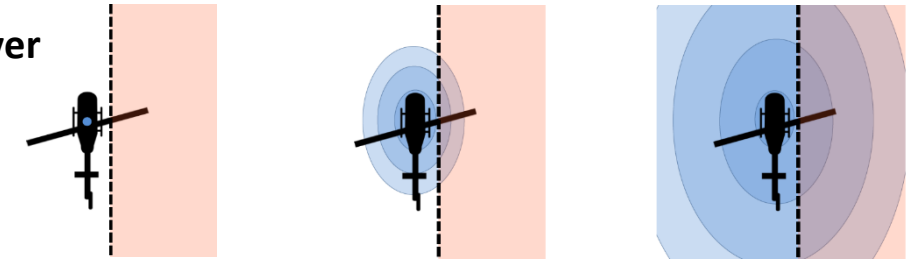
This project is funded by the EC under H2020 ICT-48

Safe Autonomous Systems

If things can go wrong they probably will!

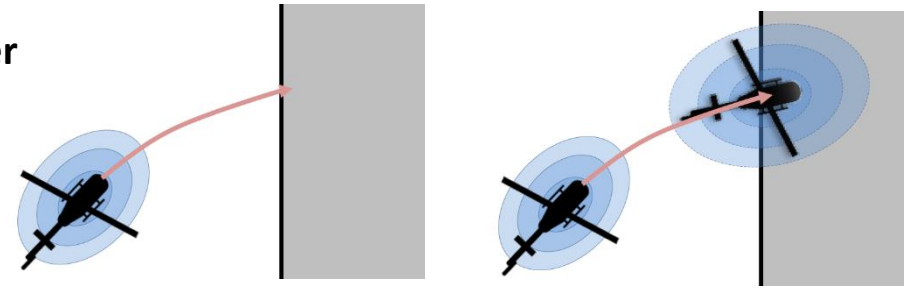
This implies the need for continual monitoring of an autonomous system and its environment in a principled, contextual, task specific manner which can be specified by the system itself!

Reasoning over Uncertainty



collision: false | Pr(collision) = 0.1 | Pr(collision) = 0.4

Reasoning over Predictions



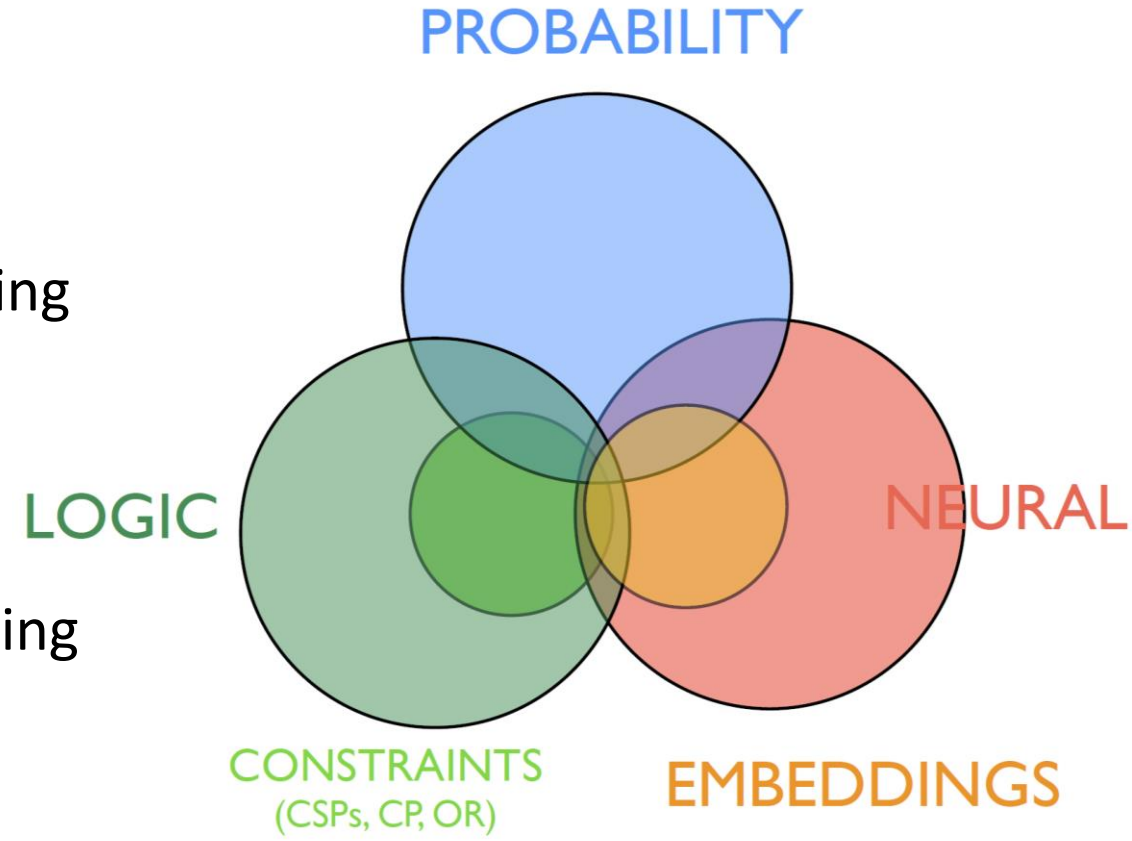
Pr(collision now) = 0.0... | Pr(collision soon) = 0.5

Incremental Reasoning in Probabilistic Signal Temporal Logic [Tiger and Heintz IJAR2020]

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

Paradigms and Representations

- Goals:
  - Integrate these paradigms
  - Integrate the involved communities
  - Covers five core different communities including
    - Deep & Probabilistic Learning
    - Neuro-Symbolic Computation (NeSy)
    - Statistical Relational AI (StarAI)
    - Constraint Programming & Machine Learning
    - Knowledge graphs for reasoning
    - And apply … in e.g. computer vision

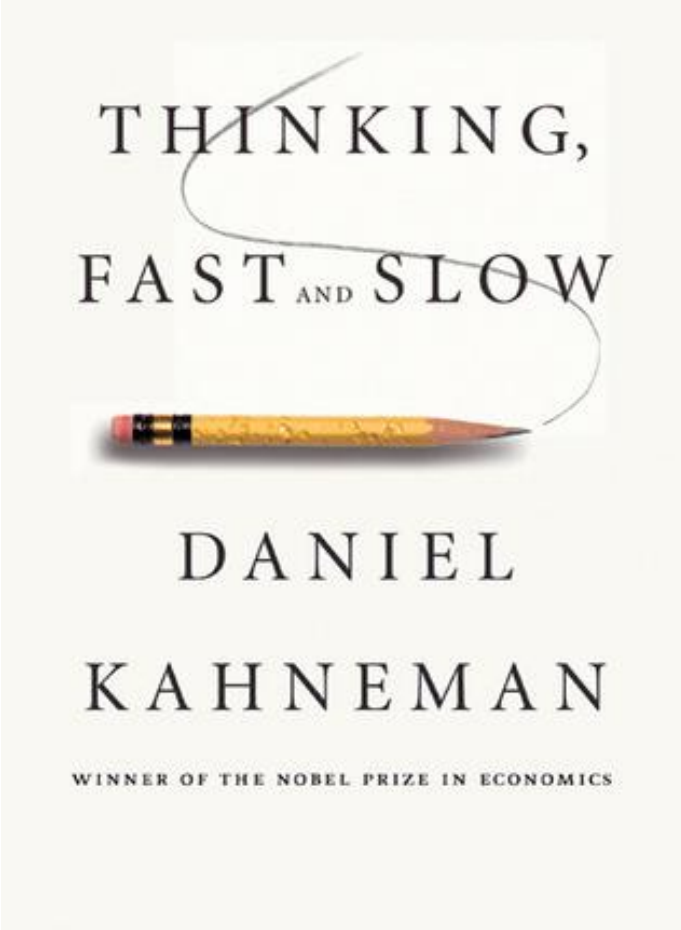


This project is funded by the EC
under H2020 ICT-48

Human and Computational Thinking

Figure 1: A Comparison of System 1 and System 2 Thinking

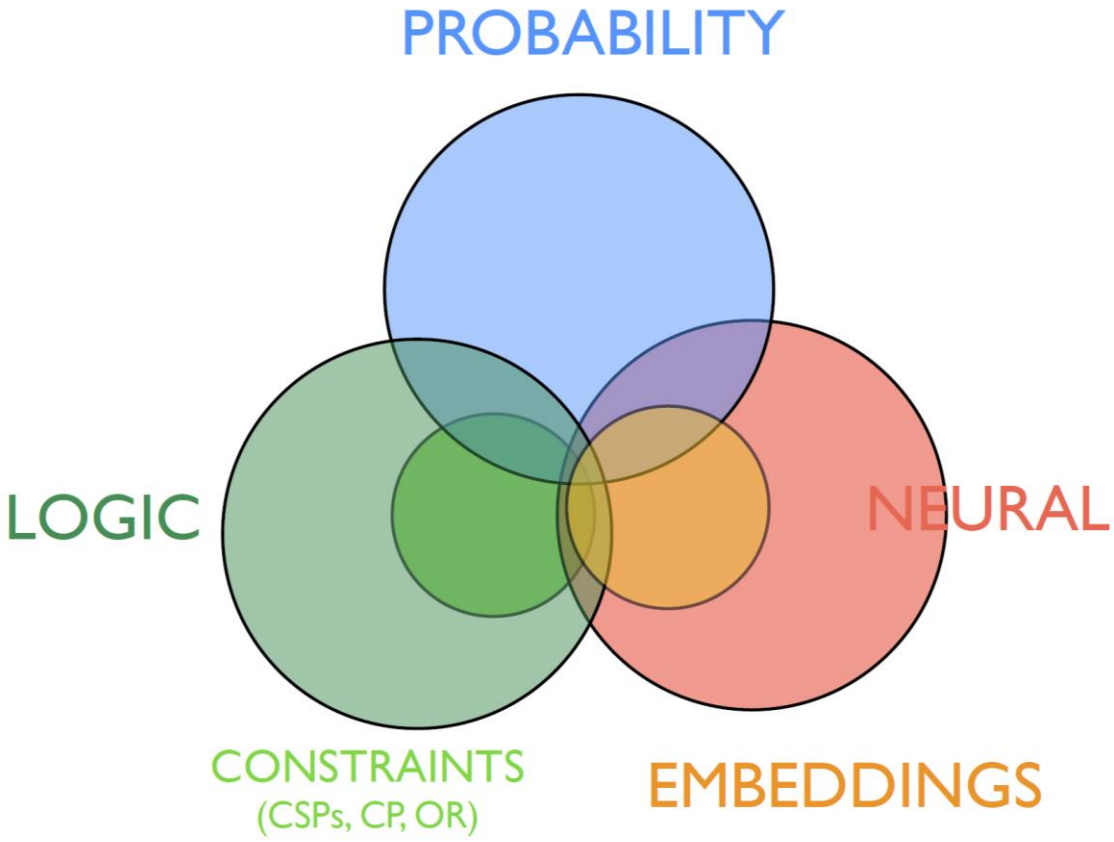
System 1 "Fast"	System 2 "Slow"
DEFINING CHARACTERISTICS Unconscious Effortless Automatic	DEFINING CHARACTERISTICS Deliberate and conscious Effortful Controlled mental process
WITHOUT self-awareness or control	WITH self-awareness or control
"What you see is all there is."	Logical and skeptical
ROLE Assesses the situation Delivers updates	ROLE Seeks new/missing information Makes decisions



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

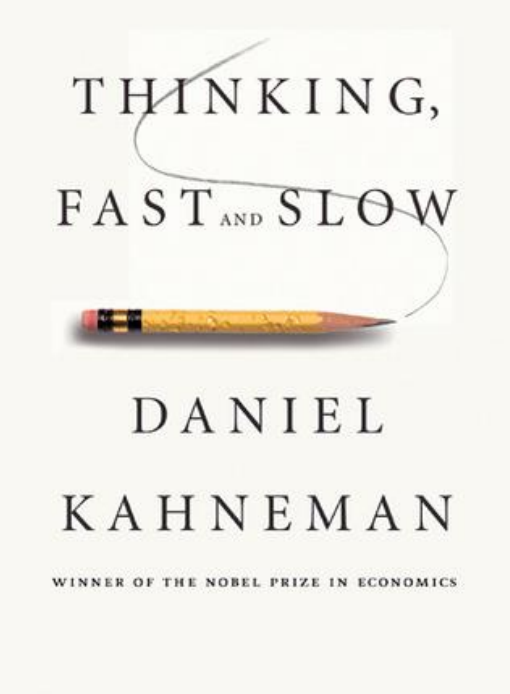
WP4: Paradigms and Representations

- **Task 4.1**: Integrated representations for learning and reasoning
- **Task 4.2:** Integrated approaches to learning and optimization
- **Task 4.3:** Learning and reasoning with embeddings, knowledge graphs, & ontologies
- **Task 4.4:** Learning and reasoning for perception, spatial reasoning, and vision



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

The Way Forward



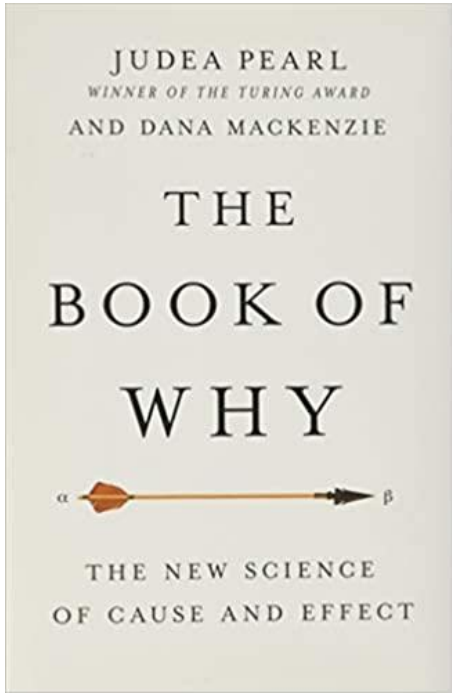
Data

Knowledge/
Assumptions

Causal Models
Combining reasoning
and learning

Explanations

Predictions

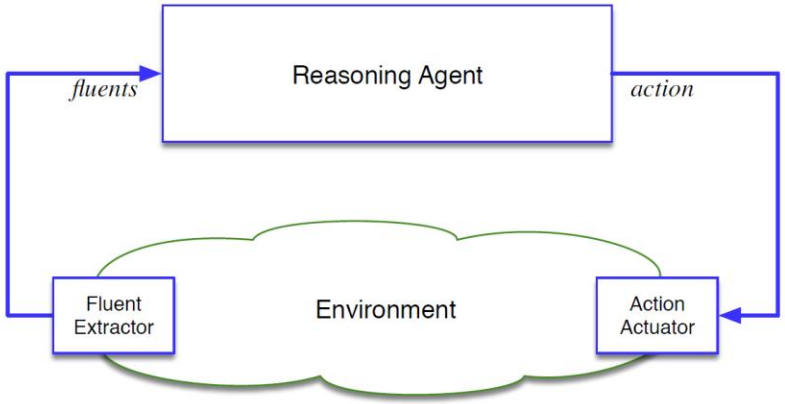


This project is funded by the EC
under H2020 ICT-48

Reasoning Agents and Learning Agents

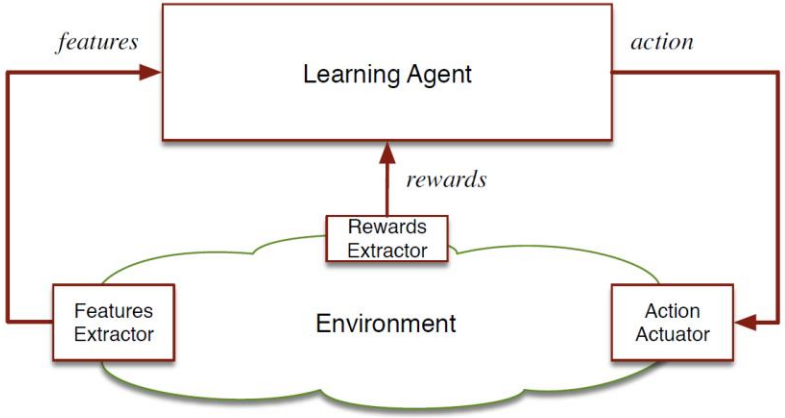
Reasoning agent:

- Senses and acts on the environment
- Has model of its environment and task
- Does Planning



Learning agent:

- Senses and acts on the environment
- Gets rewards when right
- Does Reinforcement Learning

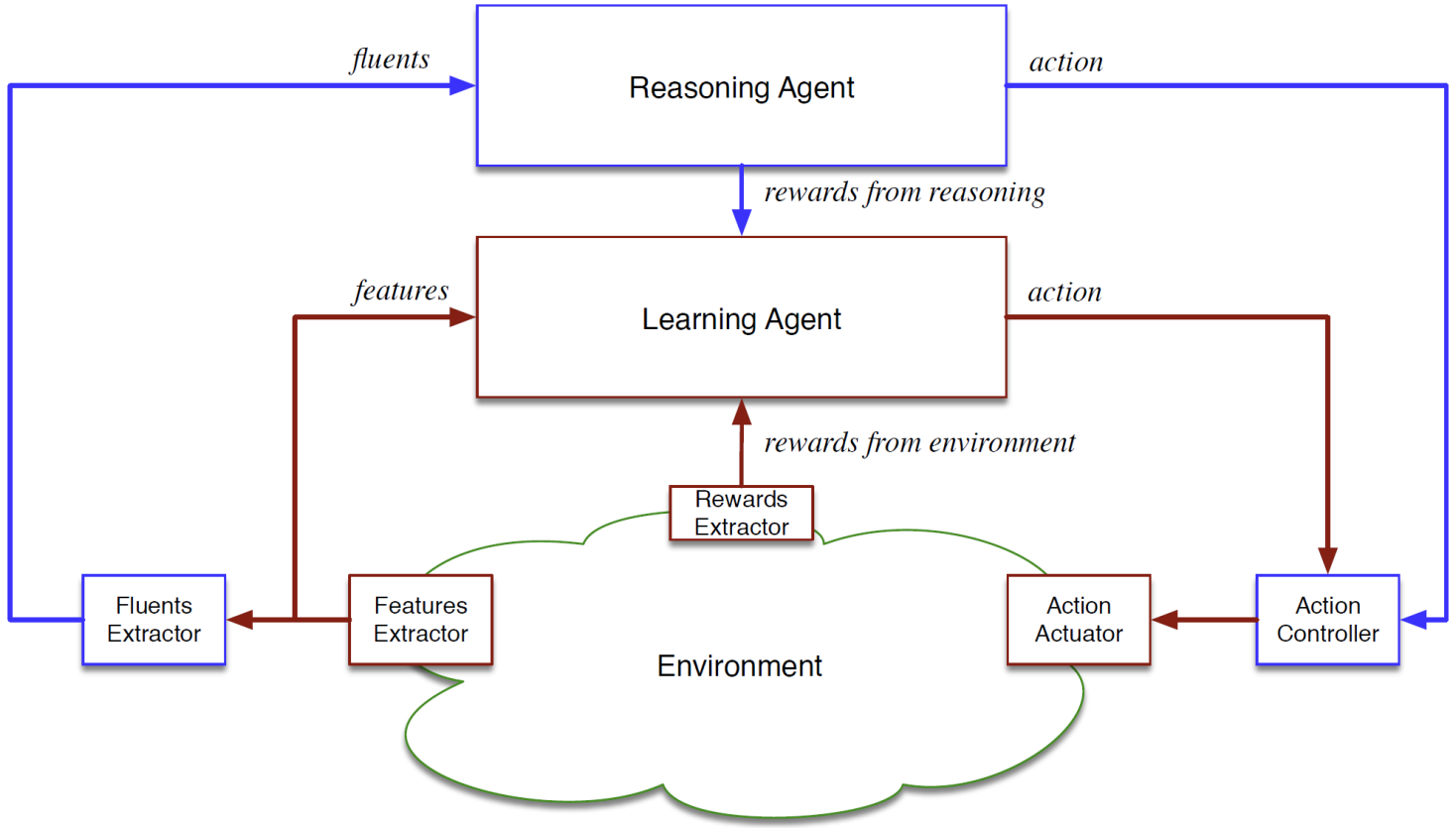


This project is funded by the EC under H2020 ICT-48

Reasoning and Learning Agents

Merging:

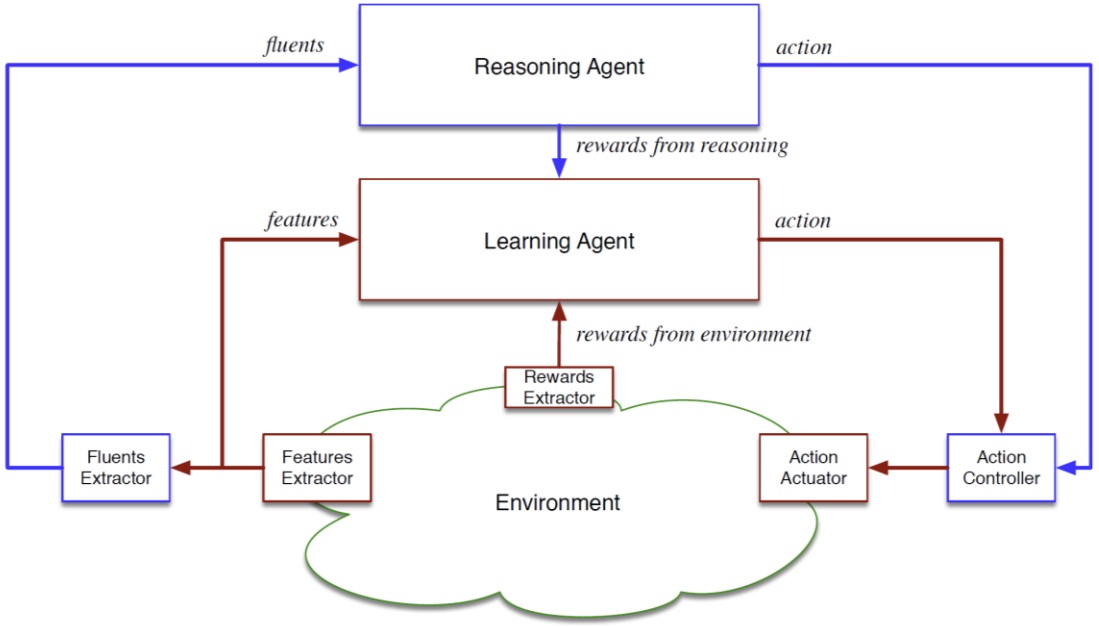
- Reasoning agent
  - E.g. reasoning in temporal logics
- Learning agent
  - E.g. doing reinforcement learning



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

WP5: Acting

- **Task 5.1:** Extended and multi-facet models of the world dynamics and tasks
- **Task 5.2:** Integrating data-based methods with model-based methods in deciding and learning how to act
- **Task 5.3:** Learning for reasoners and planners, and reasoners and planners for learning
- **Task 5.4:** 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

- **Task 6.1**: Modelling social cognition, collaboration and teamwork
  - **Integrate individual knowledge** with **knowledge available to and from other agents**
  - **Designing social AI systems**
- **Task 6.2**: 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
- **Task 6.3**: Learning from others
  - **Social learning**.
- **Task 6.4**: Emergent behaviour, agent societies and social networks
  - **Designing complex social structures, organizations and institutions**

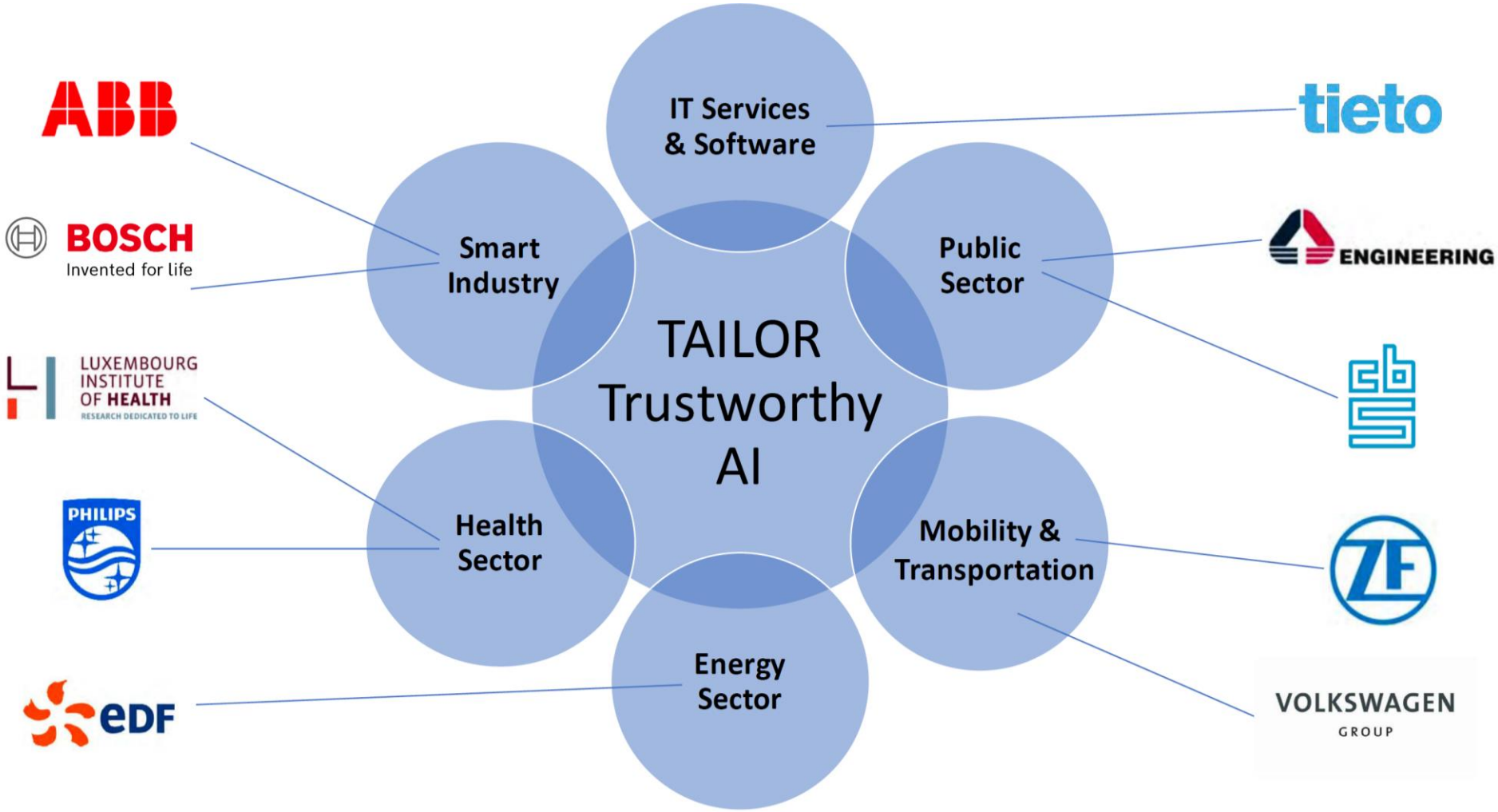
This project is funded by the EC under H2020 ICT-48

WP7: AutoAI

- Automate labour-intensive, error-prone aspects of building AI systems to make them more trustworthy and robust
- **Task 7.1:** AutoML in the wild
- **Task 7.2:** Beyond standard supervised learning
- **Task 7.3:** Self-monitoring AI Systems
- **Task 7.4:** Multi-objective AutoAI
- **Task 7.5:** Ever-learning AutoAI

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

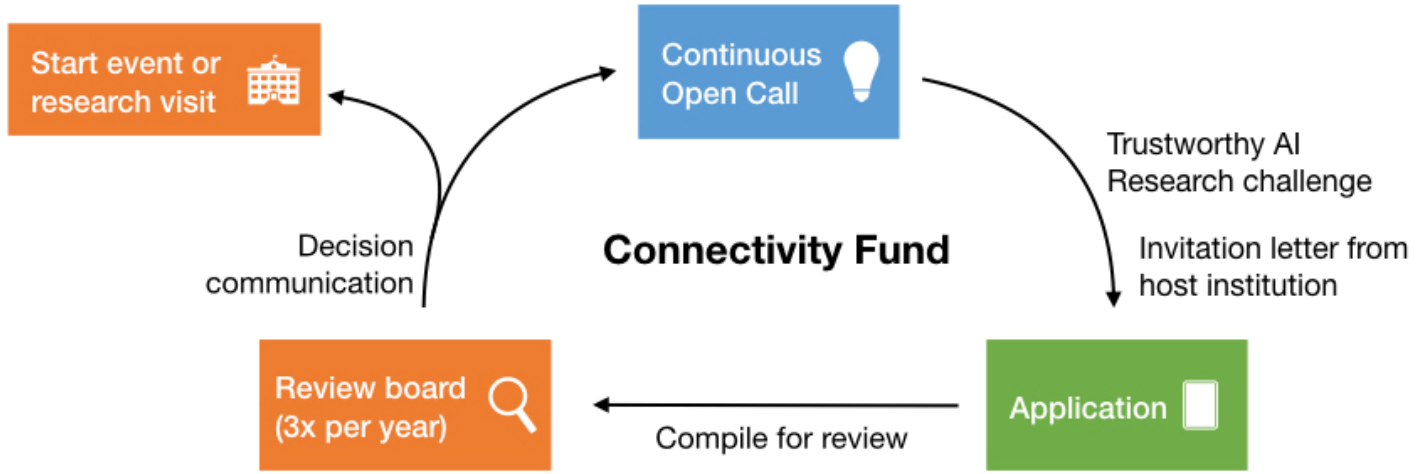
Industry Sectors and Partners



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

Connectivity Fund

- 1.5 million EUR fund, third-party funding (guest or host is non-TAILOR)
- Open call (starting late 2020), reviewed every 4 months
  - Submitted by non-TAILOR host or guest
  - Max. 60.000 EUR per visit/workshop, covers travel, housing, and sustenance
- https://tailor-eu.github.io/connectivity-fund/



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

TAILOR Objectives

O1: Establish	O2: Define and maintain	O3: Create	O4: Build	O5: Progress	O6: Increase
O1: Establish a strong pan-European network of research excellence centers on the Foundations of Trustworthy AI	O2: Define and maintain a unified strategic research and innovation roadmap for the Foundations of Trustworthy AI	O3: Create the capacity and critical mass to develop the scientific foundations for Trustworthy AI	O4: Build sustained collaborations with academic, industrial, governmental, and community stakeholders on the Foundations of Trustworthy AI	O5: Progress the Scientific State-of-the-Art for the Foundations of Trustworthy AI	O6: Increase Knowledge and Awareness of the Foundations of Trustworthy AI across Europe

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

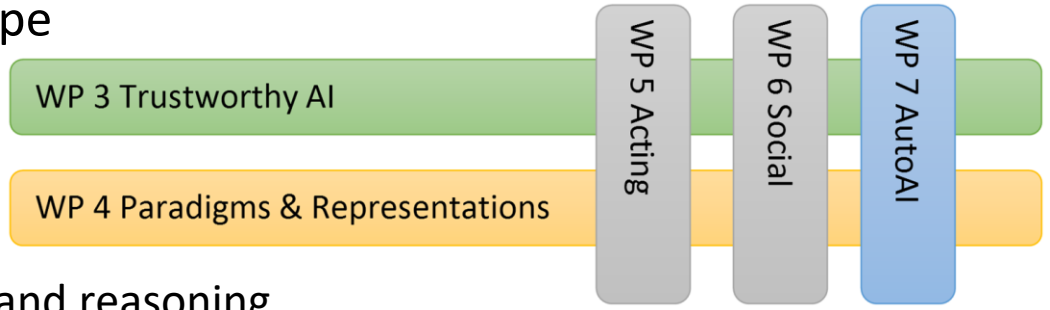
TAILOR ICT-48 Network

TAILOR https://tailor-network.eu/

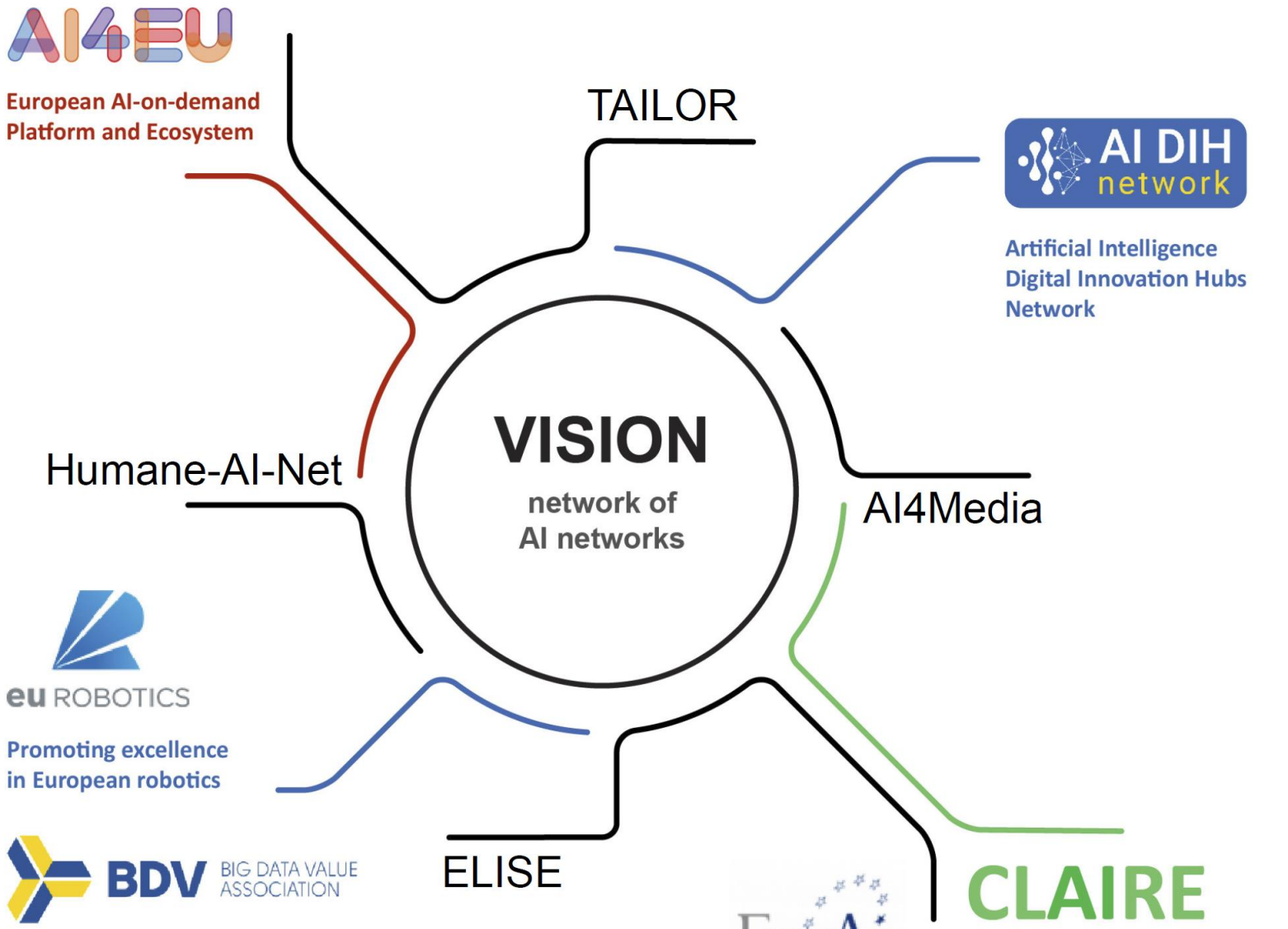
CLAIRE

*TAILOR brings together 54 leading AI research centres from **learning**, **optimisation** and **reasoning** together with major European companies representing important industry sectors into a single scientific network addressing the **scientific foundations of Trustworthy AI** to reduce the fragmentation, boost the collaboration, and increase the AI research capacity of Europe as well as attracting and retaining talents in Europe.*

- 54 research excellence centres from 20 countries across Europe coordinated by Fredrik Heintz, Linköping University, Sweden
- Four instruments
  - An ambitious research and innovation roadmap
  - Five basic research programs integrating learning, optimisation and reasoning in key areas for providing the scientific foundations for Trustworthy AI
  - A connectivity fund for active dissemination to the larger AI community
  - Network collaboration promoting research exchanges, training materials and events, and joint PhD supervision

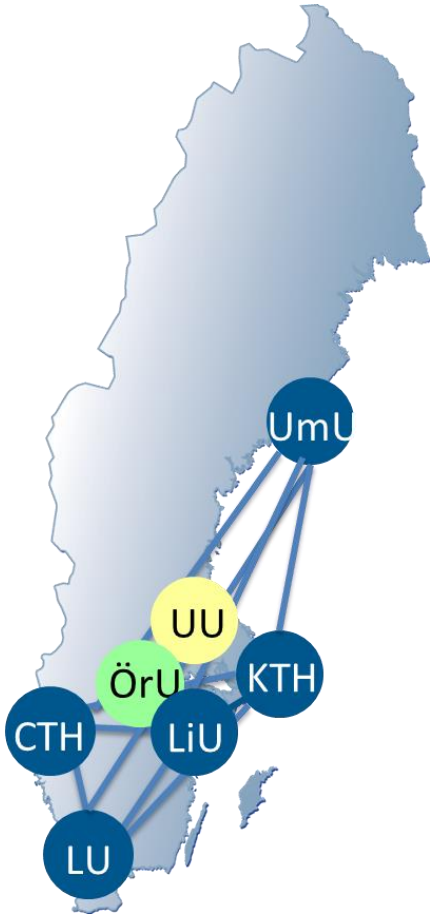


This project is funded by the EC under H2020 ICT-48



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

AI Innovation, Competence and Research in Sweden



AI SUSTAINABILITY CENTER

AI SWEDEN
AI INNOVATION of Sweden

AI Competence of Sweden

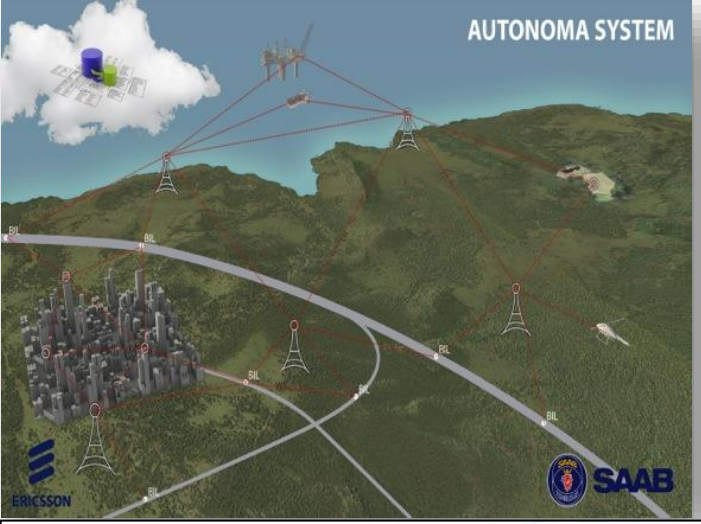
WASP
WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

RI.SE

WASP—HS

Elements of AI

CHALMERS UNIVERSITY OF TECHNOLOGY
KTH VETENSKAP OCH KONST
LINKÖPINGS UNIVERSITET
LUNDS UNIVERSITET
UMEÅ UNIVERSITET
UPPSALA UNIVERSITET
ÖREBRO UNIVERSITY



Wallenberg AI, Autonomous Systems and Software Program (WASP)

http://wasp-sweden.se/

Sweden's largest research program

14 year program 5500 MSEK (500MEUR)

Research Program

The best researchers in the field

Graduate School (600 PhDs)

Ambitious program, Industrial PhDs

Demonstrator Arenas

Demonstrations with external parties

Recruitment Program (60+ Researchers)

Internationally competitive offers

CHALMERS UNIVERSITY OF TECHNOLOGY

KTH VETENSKAP OCH KONST

LINKÖPINGS UNIVERSITET

LUNDS UNIVERSITET

UMEÅ UNIVERSITET

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

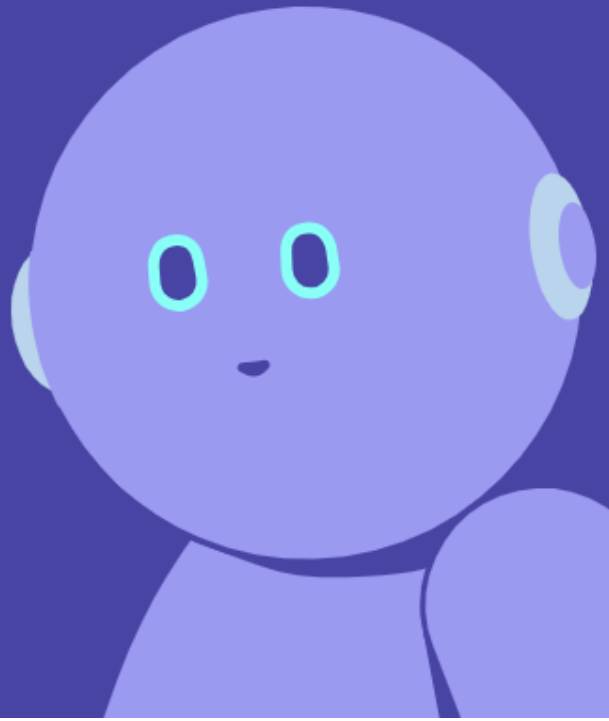
Elements of AI

Finnish →

Welcome to the Elements of Artificial Intelligence free online course

English | Start the course

Distance course at Linköping University to get 2ECTS



UNIVERSITY OF HELSINKI

Reaktor

LINKÖPINGS UNIVERSITET

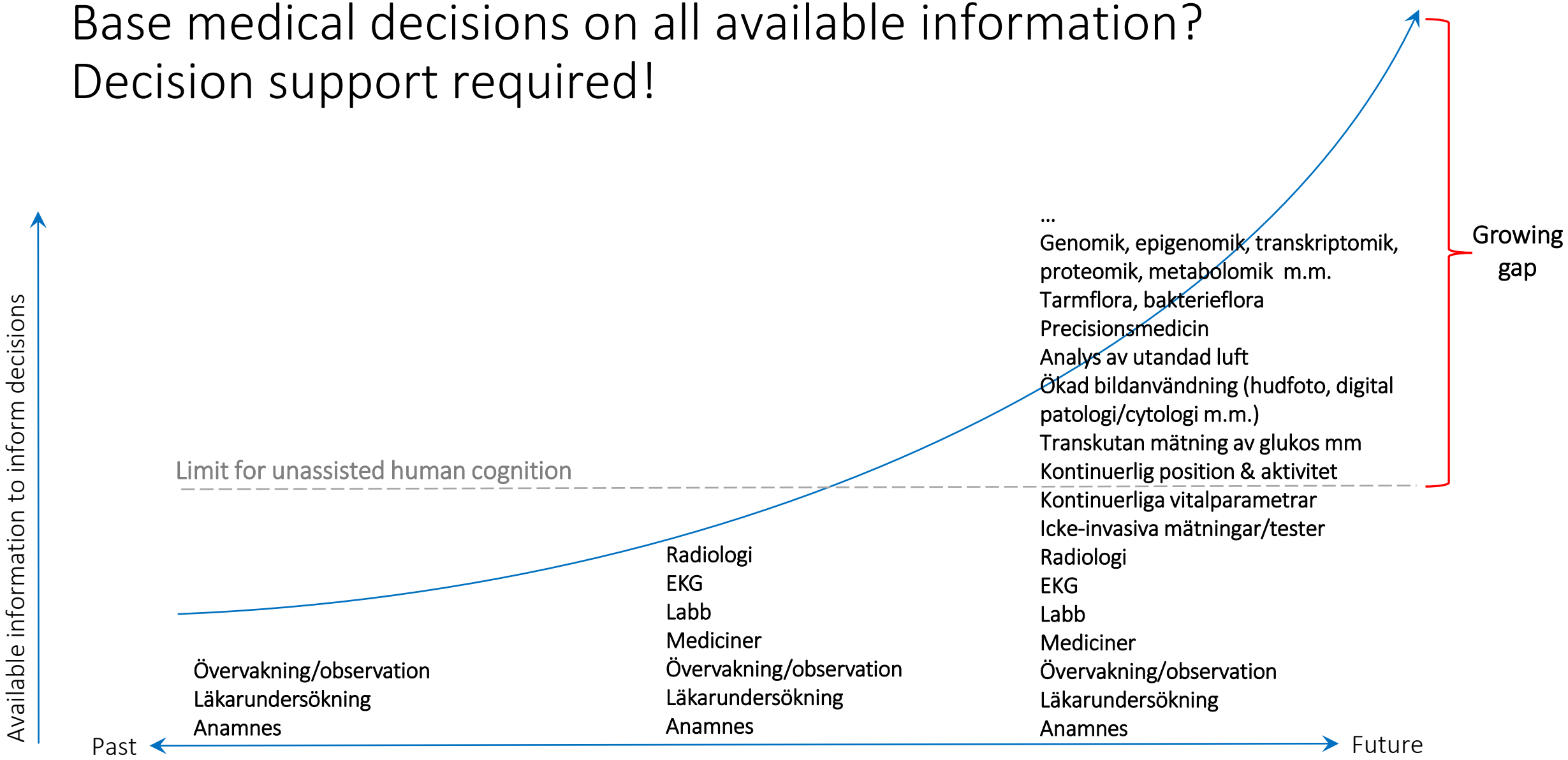
AI SWEDEN

AI COMPETENCE FOR SWEDEN

https://www.elementsofai.se/

Swedish launch funded by VINNOVA Sweden's Innovation Agency

Base medical decisions on all available information? Decision support required!



LINKÖPINGS UNIVERSITET

Illustration gjord av Erik Sundvall, Region Östergötland & LiU. Delvis baserad på: Evidence-Based Medicine and the Changing Nature of Healthcare: Workshop Summary (IOM Roundtable on Evidence-Based Medicine) Mark B. McClellan, Michael McGinnis, Elizabeth G. Nabel, and LeighAnne M. Olsen, Institute of Medicine. ISBN: 0-309-11370-9 https://www.nap.edu/catalog/12041/evidence-based-medicine-and-the-changing-nature-of-health-care Fig 5-1. page 116

Most Important Aspects for Europe to Lead?

- Provide **dedicated, significant and long-term research funding** for both **fundamental** and **purpose-driven research on AI** to promote AI that is trustworthy and to address relevant scientific, ethical, sociocultural and industrial challenges.
- Create incentives and support for **interdisciplinary** and **multi-stakeholder research** for example through large-scale challenge-driven research missions.
- Simplify and **streamline** the structure of **research funding instruments**.
- Create the proposed **lighthouse centre** in a way that effectively achieves **critical mass, synergy, and cohesion** across the European AI ecosystem.
- Invest both in **up-/reskilling** and in **basic education** related to AI.
- Establish a **clear strategy** for coordinating and structuring an **AI-based innovation ecosystem** across Europe.
- Focus "AI made in Europe" on "**AI for Good**" and "**AI for All**".
- Invest in promoting **broader awareness** of AI **in society**.


Respect for human autonomy


Prevention of harm


Fairness


Explicability

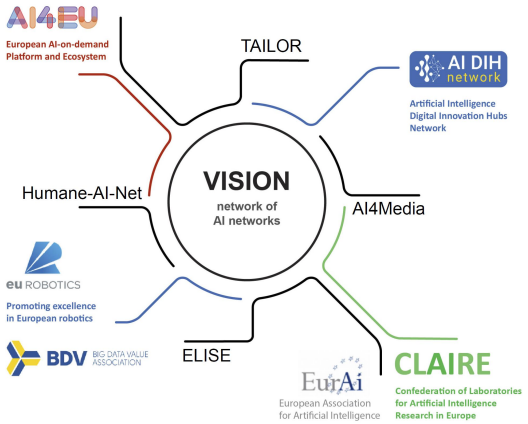
LINKÖPINGS UNIVERSITET

Fredrik Heintz, Linköping University, Sweden
fredrik.heintz@liu.se; @FredrikHeintz

TAILOR

Take Away Message

- AI is about understanding intelligence and develop systems that exhibit intelligent behavior.
- AI will affect all aspects of our society. **Trust is essential!**
- To be **trustworthy** an **AI-system** should be **legal**, **ethical** and **robust**.
- Europe has **many initiatives** in the area, but **more** is needed.
- Several important research challenges remain such as
  - safety/robustness,
  - explainability/interpretability,
  - fairness/equity/justice, and
  - governance/accountability
- Very active and interdisciplinary research problems that are still mostly unsolved.
- **The TAILOR project is committed to develop the scientific foundations for Trustworthy AI**
- **Will most likely require integrating model-free data-driven learning approaches with model-based knowledge-driven reasoning approaches**



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