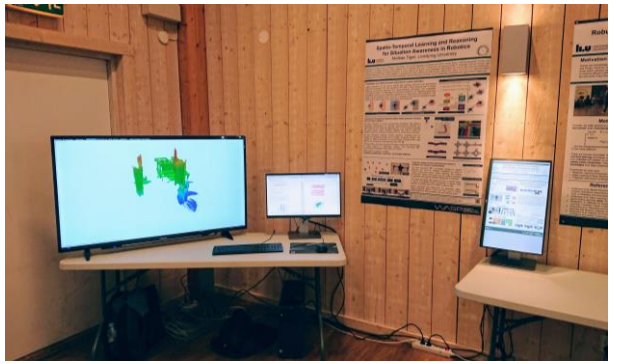
Towards Trustworthy AI
– Integrating Reasoning and Learning

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

LINKÖPINGS UNIVERSITET

20+ Years of Experience Building Fielded AI Systems



RoboCup 2000-2017

UAS Research 2000-

WARA PS 2017-

Collaborative Unmanned Aircraft Systems

A principled approach to building collaborative intelligent autonomous systems for complex missions.

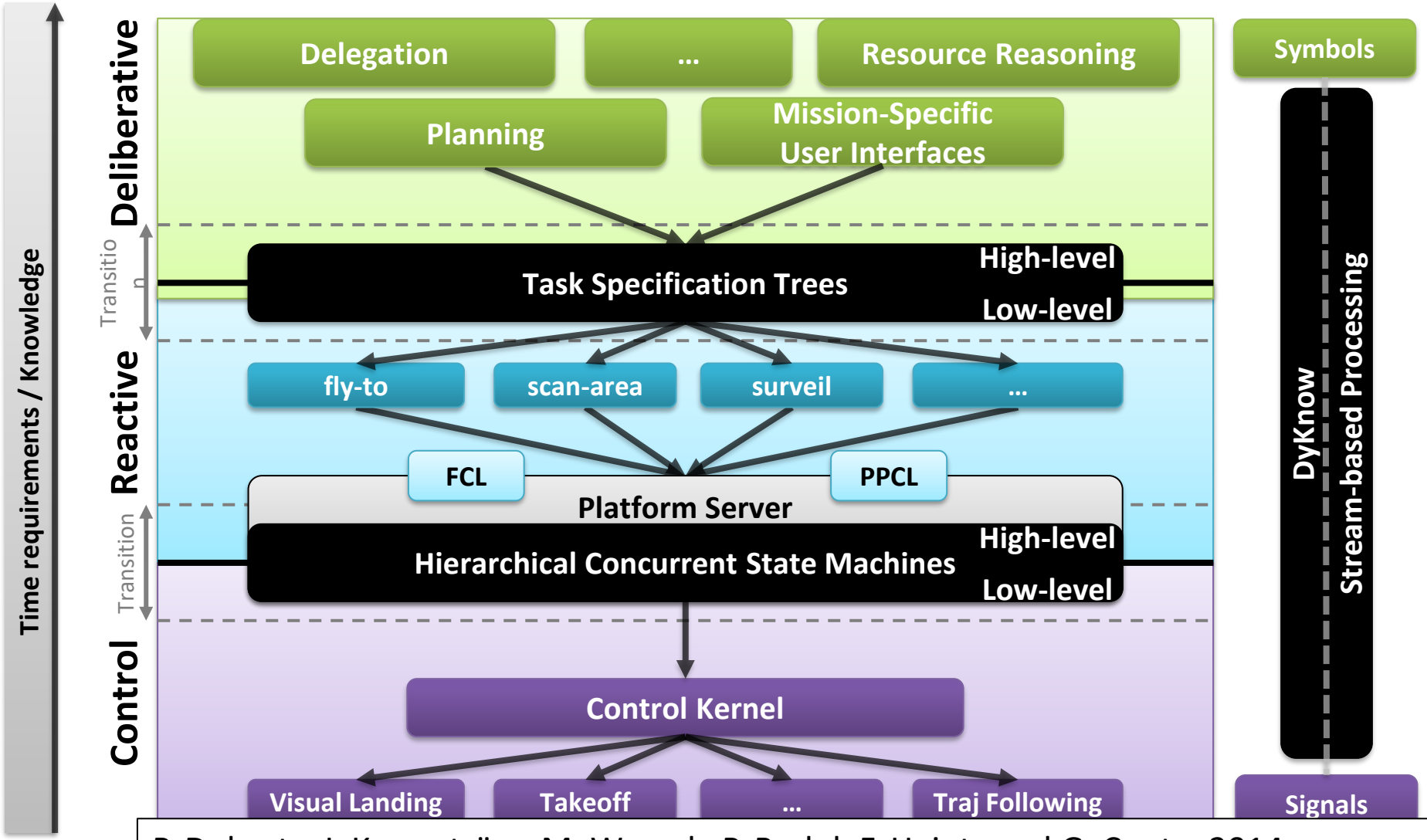
Autonomous Systems at AIICS, Linköping University

Micro UAVs
weight < 500 g,
diameter < 50 cm

Yamaha RMAX
weight 95 kg,
length 3.6 m

LinkQuad weight ~1 kg, diameter ~70cm

HDRC3: A Distributed Hybrid Deliberative/Reactive Architecture for Autonomous Systems

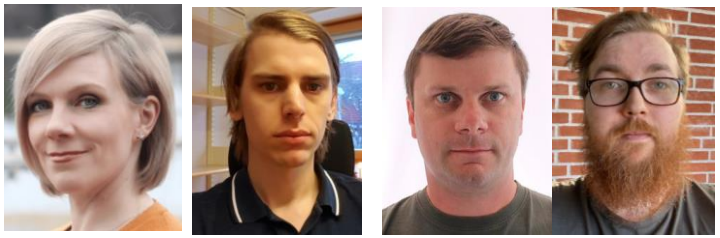


P. Doherty, J. Kvarnström, M. Wzorek, P. Rudol, F. Heintz and G. Conte. 2014.
HDRC3 - A Distributed Hybrid Deliberative/Reactive Architecture for Unmanned Aircraft Systems.
In K. Valavanis, G. Vachtsevanos, editors, Handbook of Unmanned Aerial Vehicles, pages 849–952.

LINKÖPINGS UNIVERSITET

Reasoning and Learning Lab

1 Adj prof 6+ PhD students

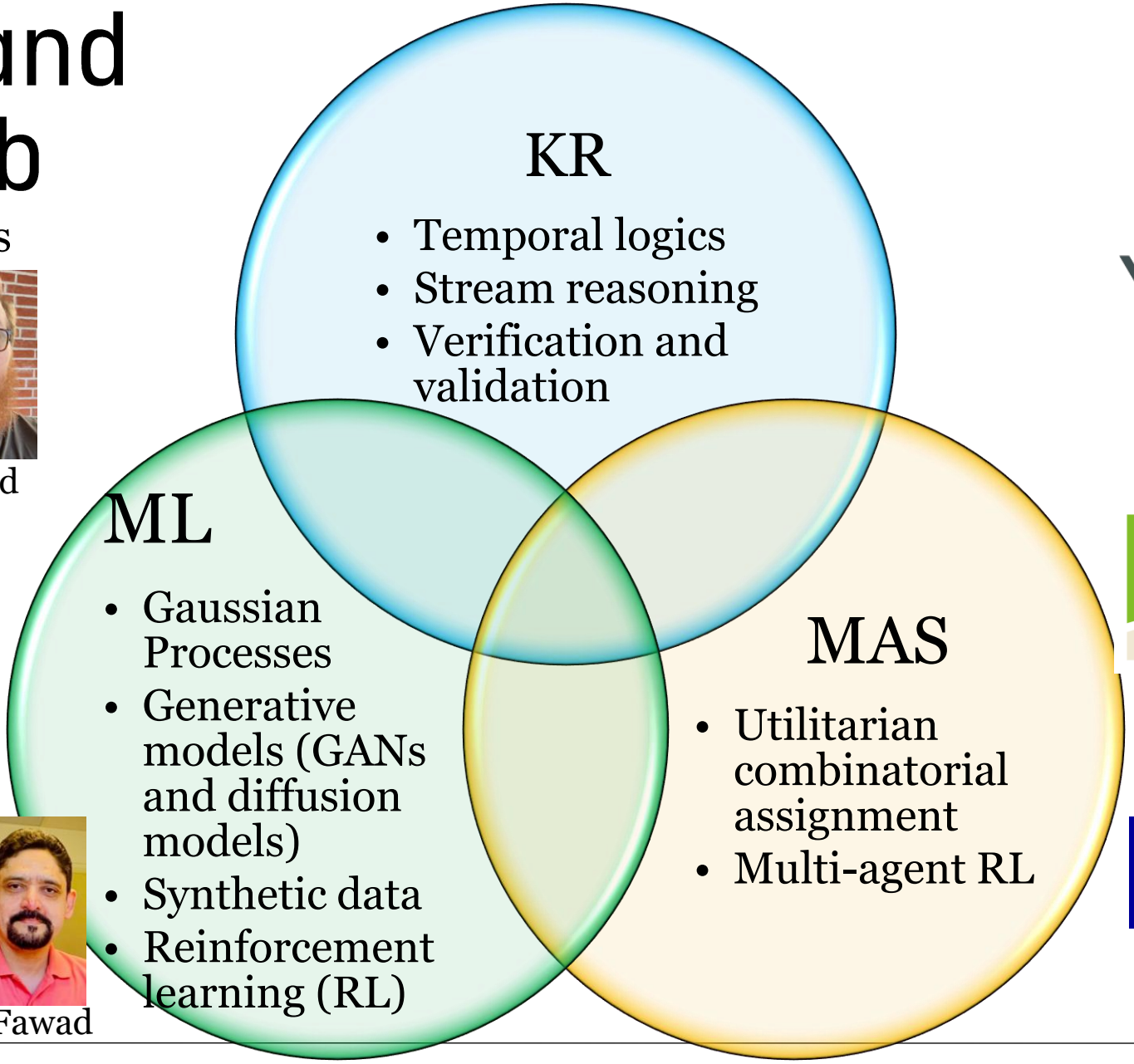


Linda, Fredrik, Johan, David

Fahim, Dennis, Mohsen

4+ Postdocs, 1 Coordinator

Resmi, Mattias, Katerina, Daniel, Fawad



Knut and Alice Wallenberg Foundation

WASP

STIFTELSEN MARCUS OCH AMALIA WALLENBERGS MINNESFOND

VINNOVA

ELLIIT

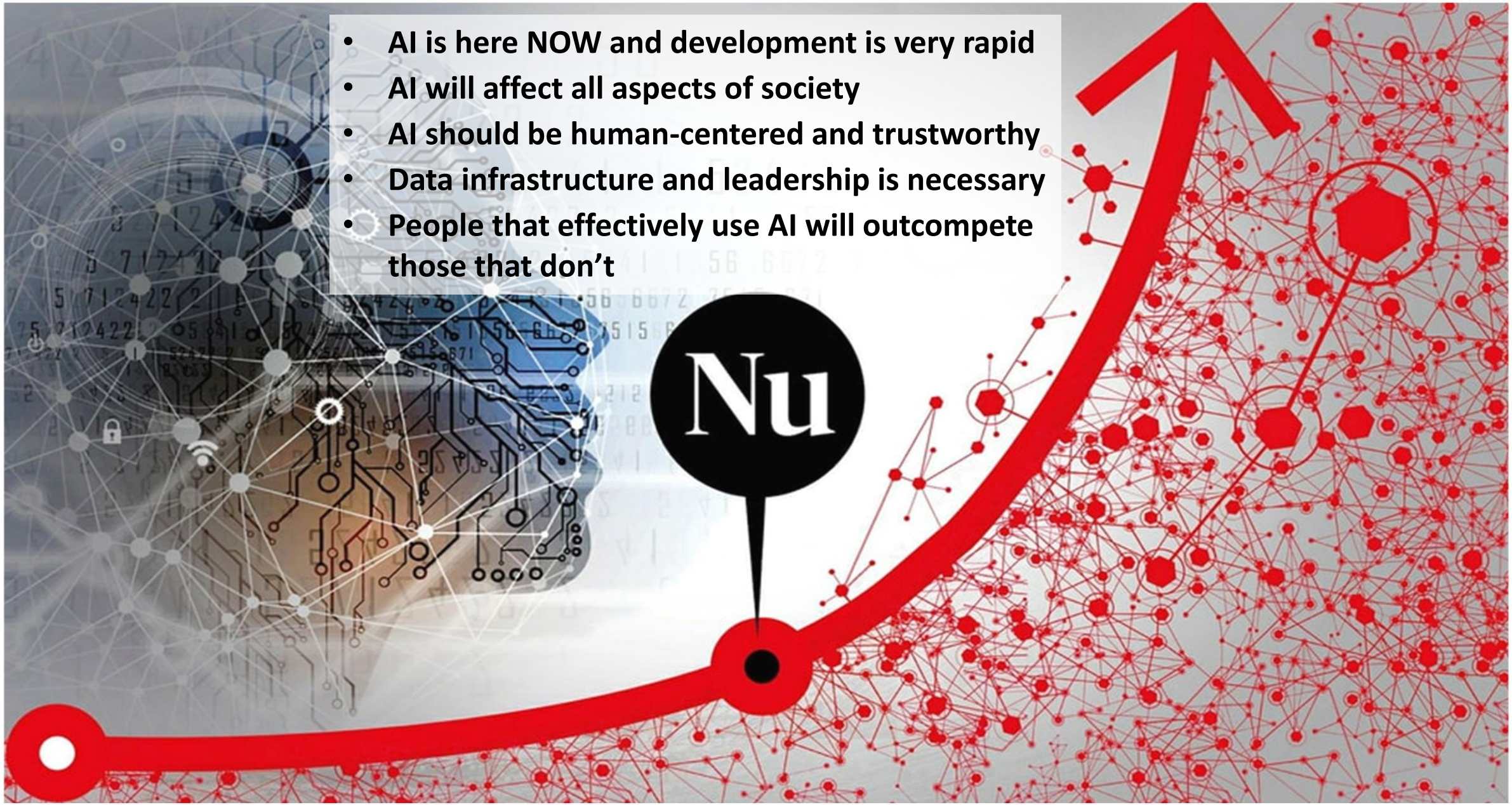
TAILOR

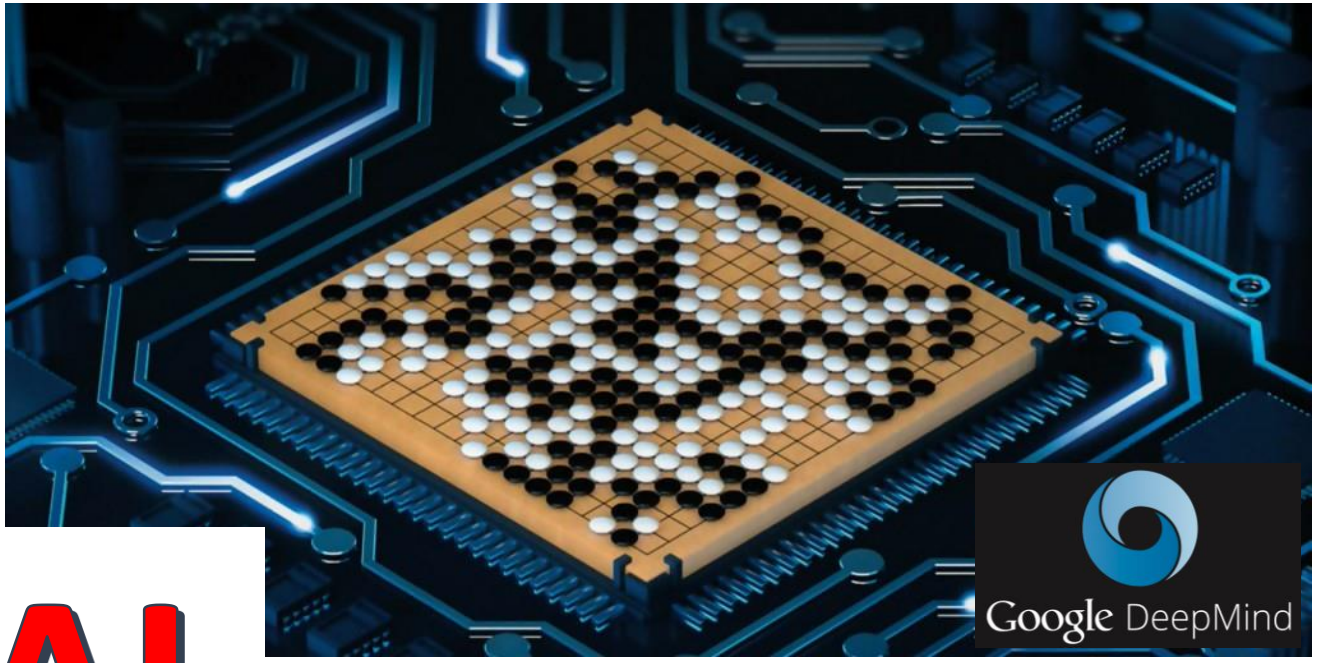
WASP—HS

LINKÖPING UNIVERSITY

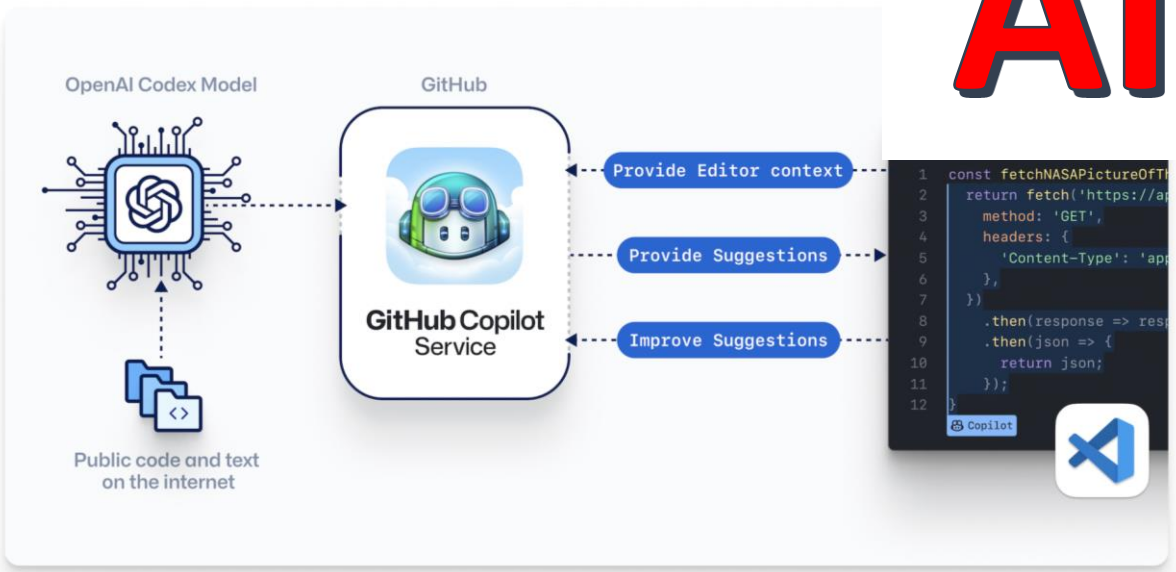
DAGENS NYHETER.

- AI is here NOW and development is very rapid
- AI will affect all aspects of society
- AI should be human-centered and trustworthy
- Data infrastructure and leadership is necessary
- People that effectively use AI will outcompete those that don’t

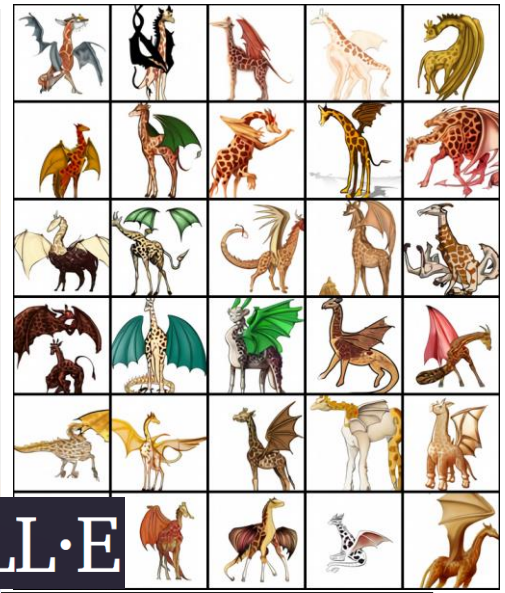



Google DeepMind

AI

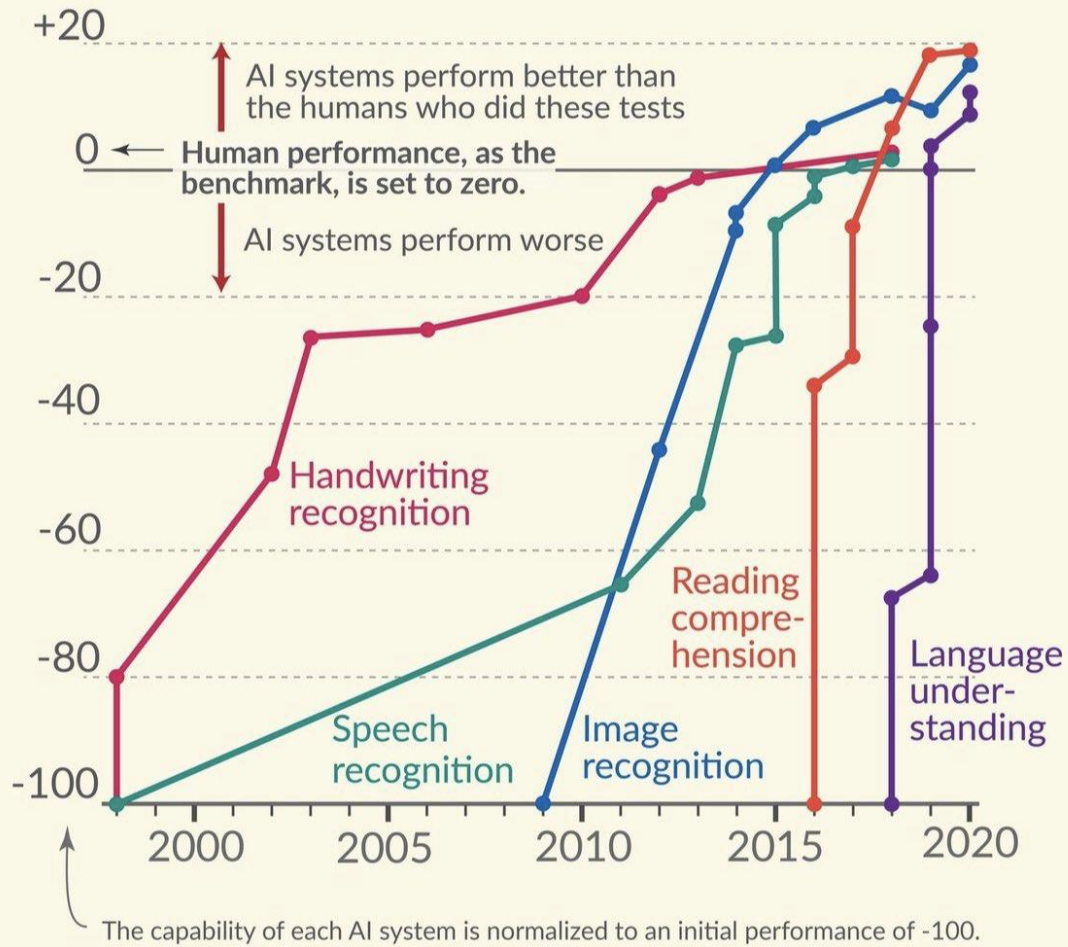

OpenAI Codex Model
GitHub
GitHub Copilot Service
Provide Editor context
Provide Suggestions
Improve Suggestions
Public code and text on the internet
Copilot


DALL·E



Language and image recognition capabilities of AI systems have improved rapidly

Test scores of the AI relative to human performance



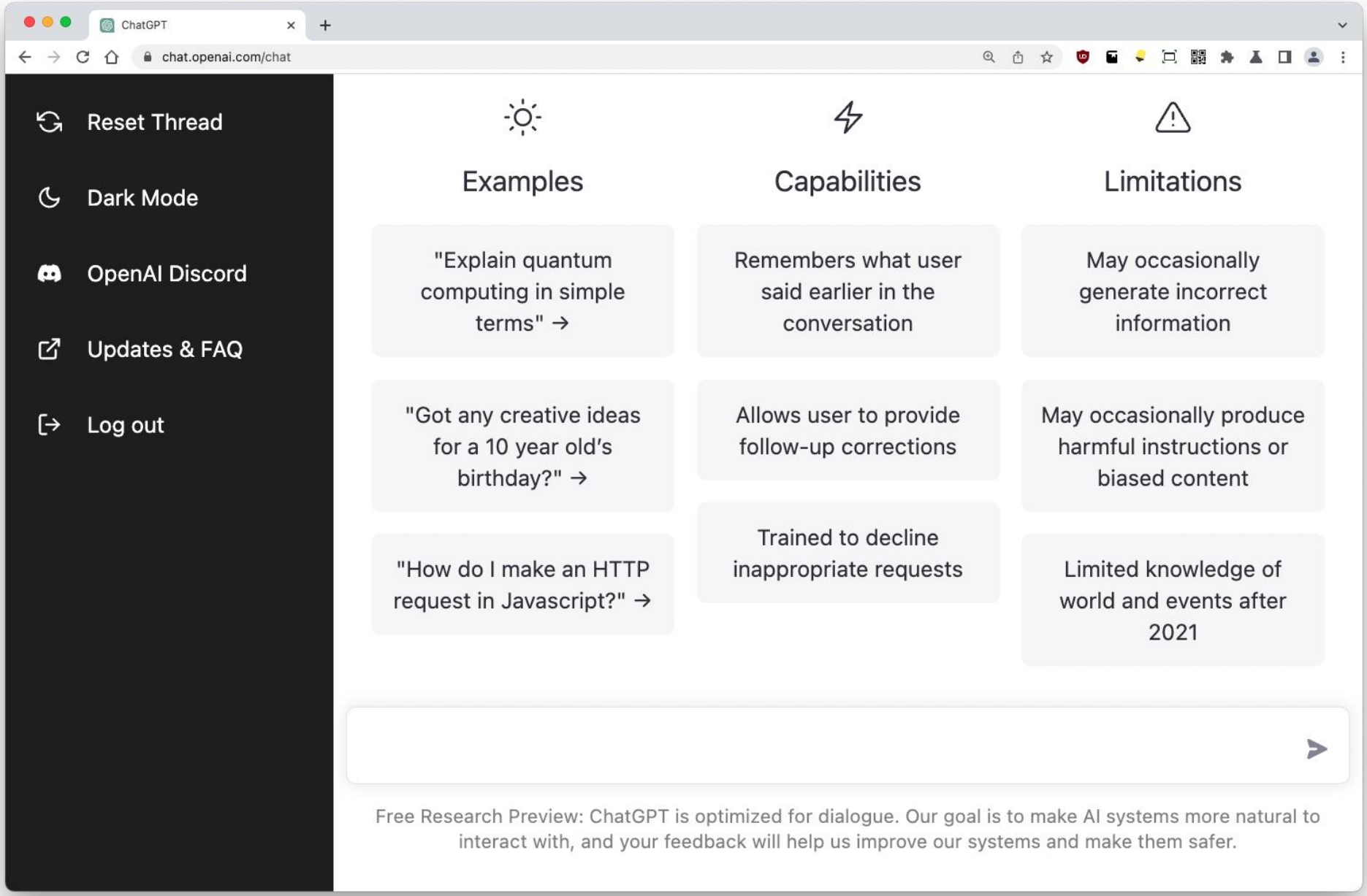
Source:
Kiela et al. (2021) Dynabench: Rethinking Benchmarking in NLP

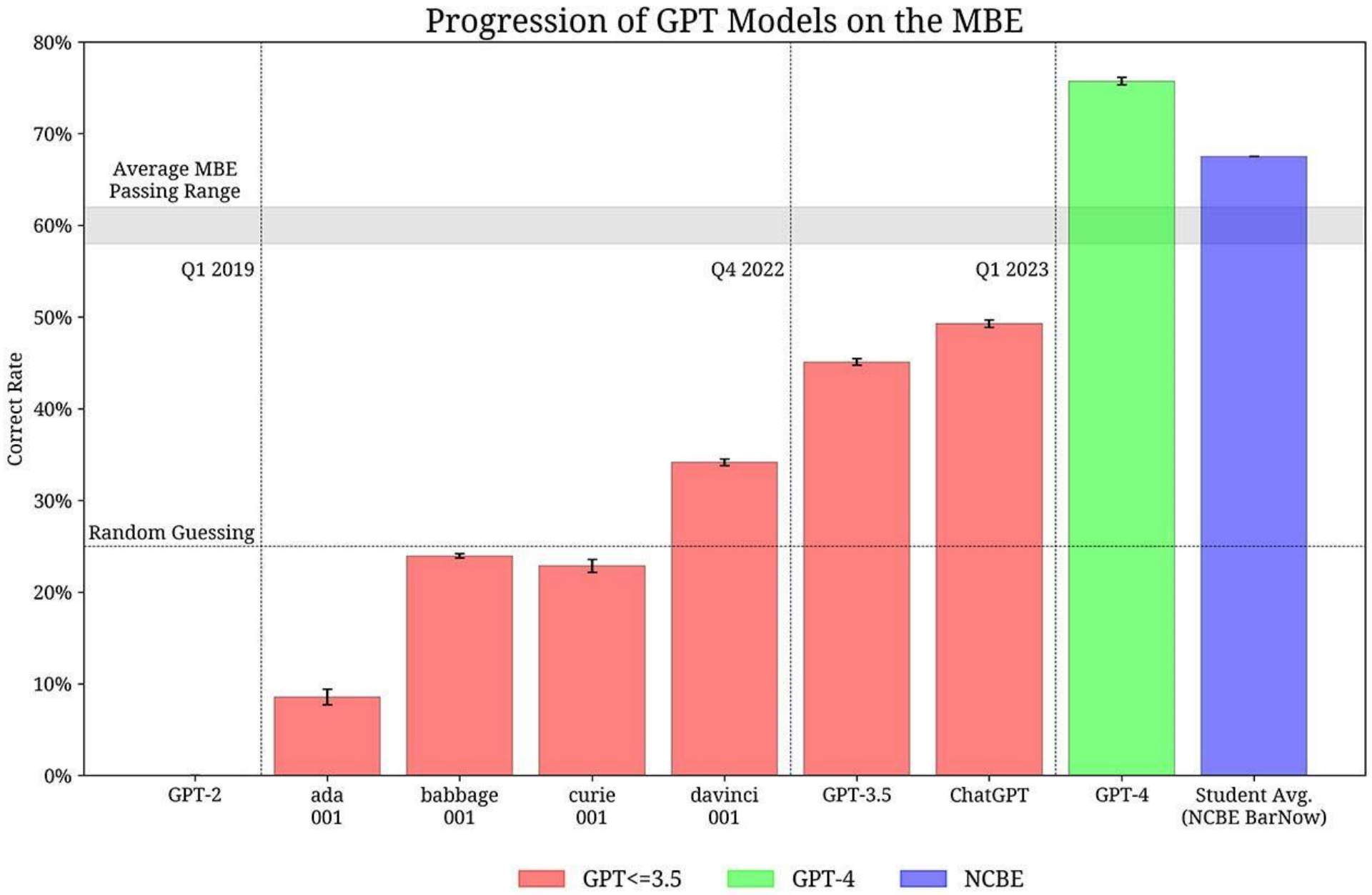
OurWorldInData.org/artificial-intelligence • CC BY

Our World in Data

“Weak human + machine + superior process was greater than a strong computer and, remarkably, greater than a strong human + machine with inferior process.”

Garry Kasparov


ChatGPT
chat.openai.com/chat
Reset Thread
Dark Mode
OpenAI Discord
Updates & FAQ
Log out
Examples
Capabilities
Limitations
"Explain quantum computing in simple terms" →
"Got any creative ideas for a 10 year old's birthday?" →
"How do I make an HTTP request in Javascript?" →
Remembers what user said earlier in the conversation
Allows user to provide follow-up corrections
Trained to decline inappropriate requests
May occasionally generate incorrect information
May occasionally produce harmful instructions or biased content
Limited knowledge of world and events after 2021
Free Research Preview: ChatGPT is optimized for dialogue. Our goal is to make AI systems more natural to interact with, and your feedback will help us improve our systems and make them safer.



Katz, D., Bommarito, M., Gao, S. and Arredondo, P. *GPT-4 Passes the Bar Exam* (March 15, 2023). https://ssrn.com/abstract=4389233

How good are LLMs?

Exploring the MIT Mathematics and EECS Curriculum Using Large Language Models

Sarah J. Zhang* MIT sjzhang@mit.edu

Sam Florin* MIT sflorin@mit.edu

Ariel N. Lee Boston University ariellee@bu.edu

Eamon Niknafs Boston University en@bu.edu

Andrei Marginean MIT atmargi@mit.edu

Annie Wang MIT annewang@mit.edu

Keith Tyser Boston University ktyser@bu.edu

Zad Chin Harvard University zadchin@college.harvard.edu

Yann Hicke Cornell University ylh8@cornell.edu

Nikhil Singh MIT nsingh1@mit.edu

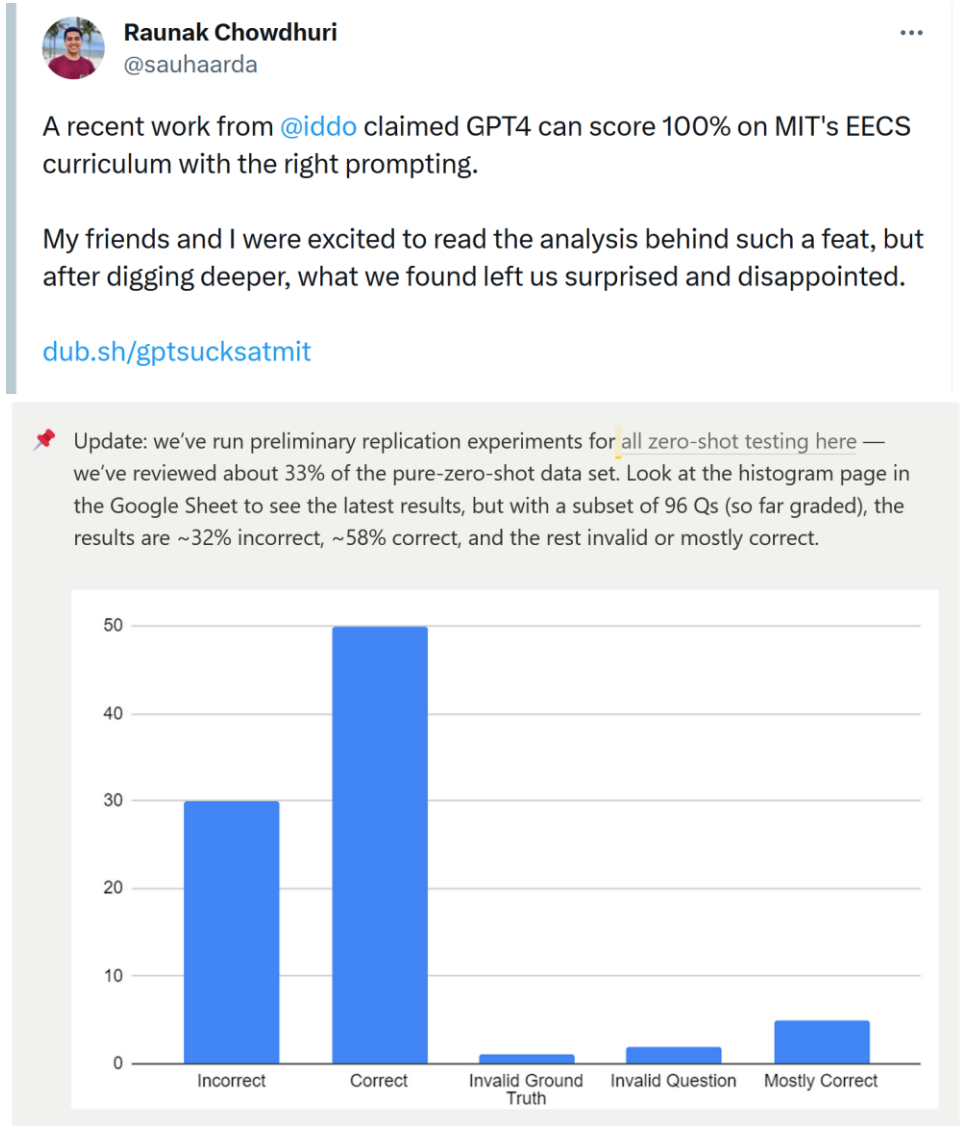
Madeleine Udell Stanford University udell@stanford.edu

Yoon Kim MIT yoonkim@mit.edu

Tonio Buonassisi MIT buonassi@mit.edu

Armando Solar-Lezama MIT asolar@csail.mit.edu

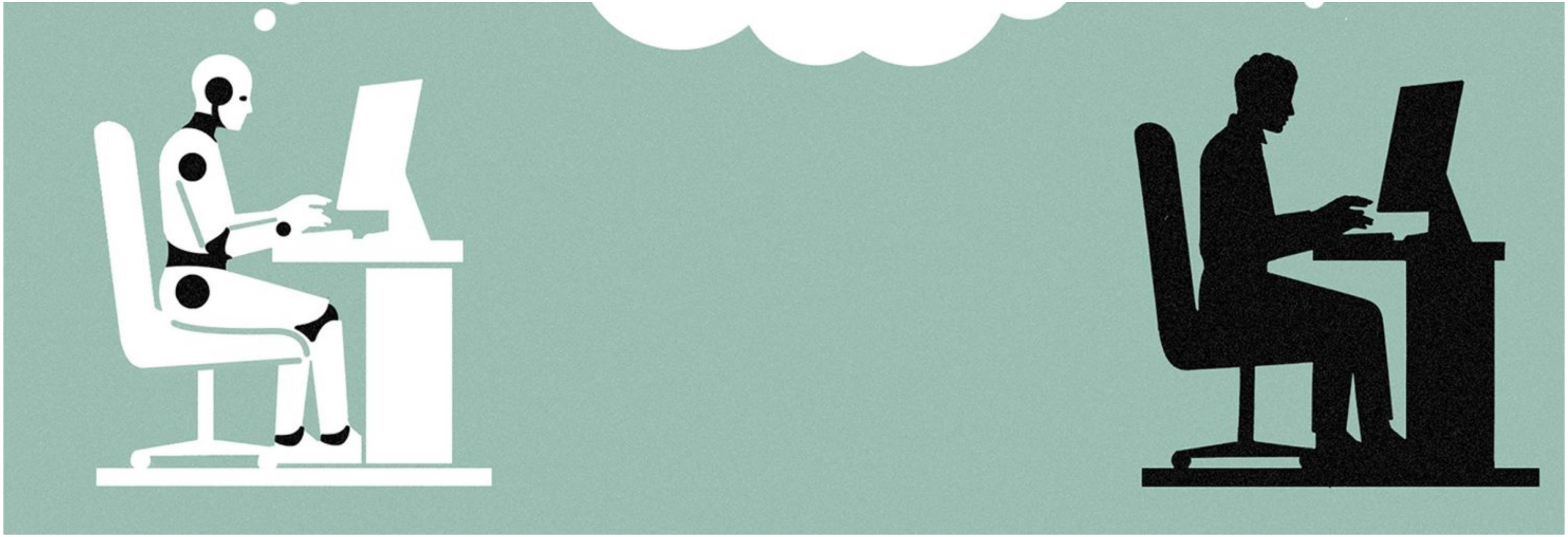
Iddo Drori MIT, Columbia University, Boston University idrori@csail.mit.edu



https://arxiv.org/abs/2306.08997

LINKÖPING UNIVERSITY

WASP ED
WALLENBERG AI AND TRANSFORMATIVE TECHNOLOGIES EDUCATION DEVELOPMENT PROGRAM



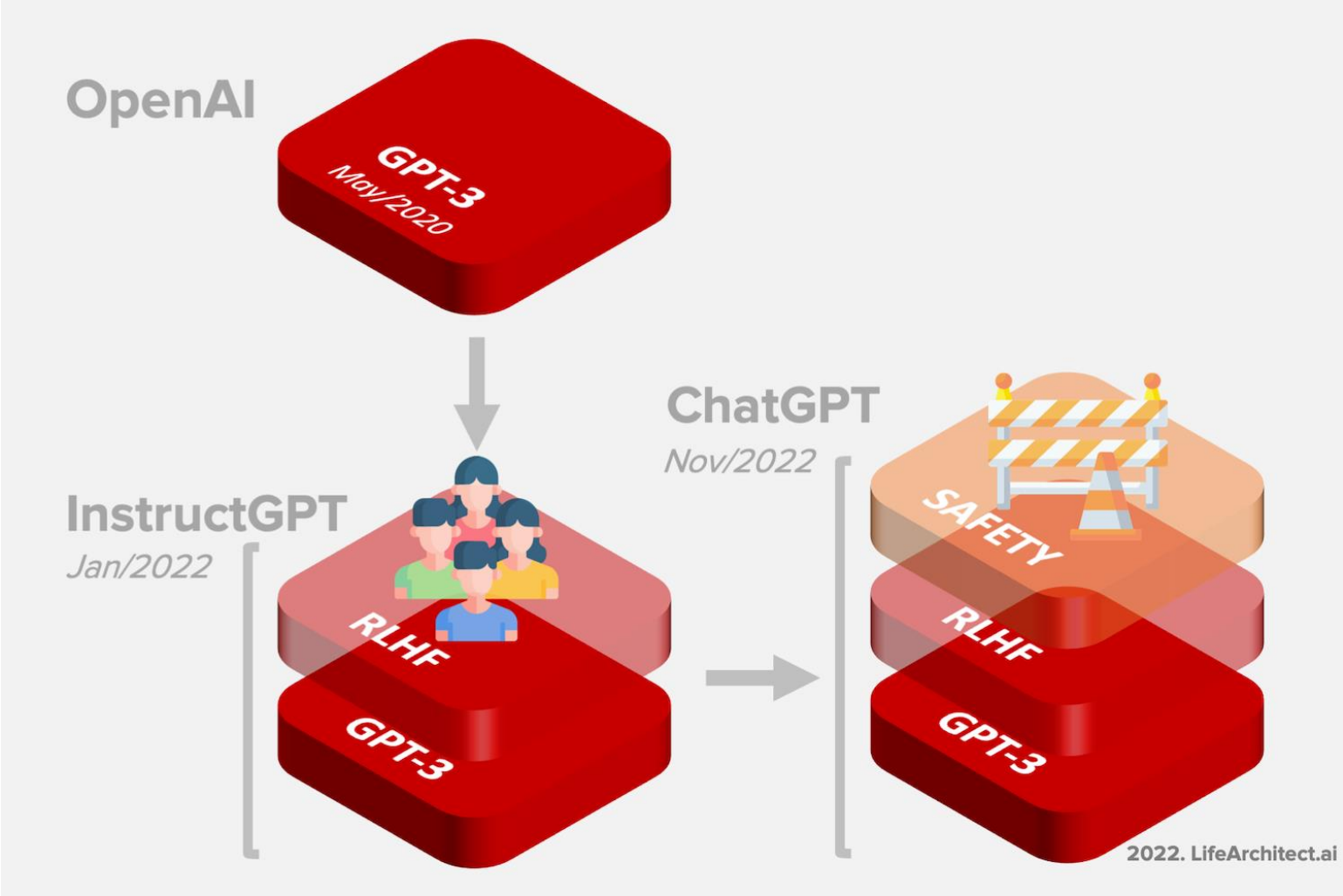
Researchers pitted Wharton students against ChatGPT and a version of ChatGPT trained with examples to see which came up with better product ideas. DAN PAGE

TECHNOLOGY | ARTIFICIAL INTELLIGENCE

M.B.A. Students vs. ChatGPT: Who Comes Up With More Innovative Ideas?

We put humans and AI to the test. The results weren't even close.

How does ChatGPT Work?

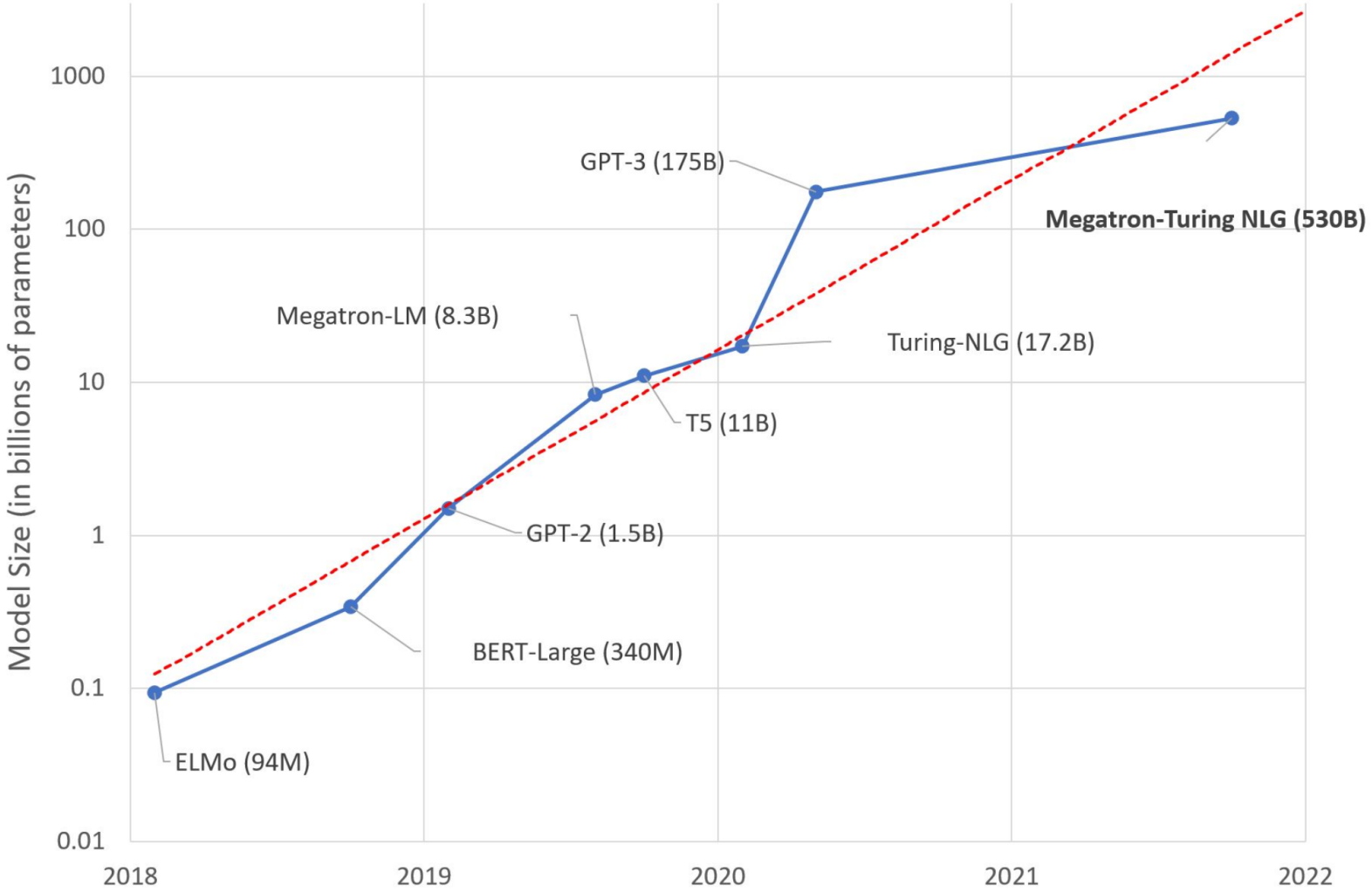


https://lifearchitect.ai/chatgpt/

Can you Trust ChatGPT? No!

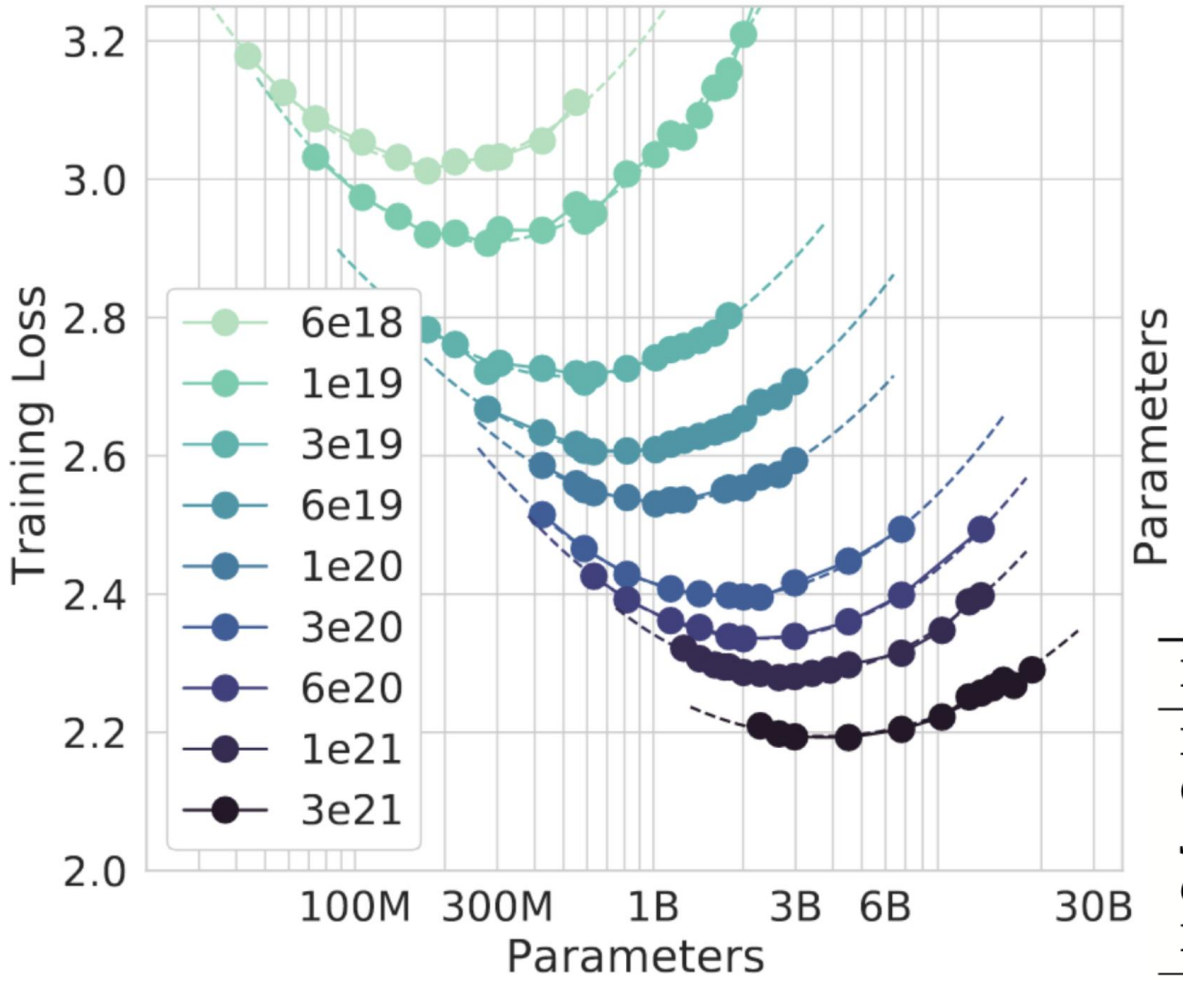
- Very limited information about the training data
- It hallucinates
- Even when there are references these may be false or not applicable
- Cannot count or draw logical conclusions
- *but, ChatGPT is still useful!*

Large Language Models – Scaling over Time



https://huggingface.co/blog/large-language-models

Large Language Models Scaling Laws



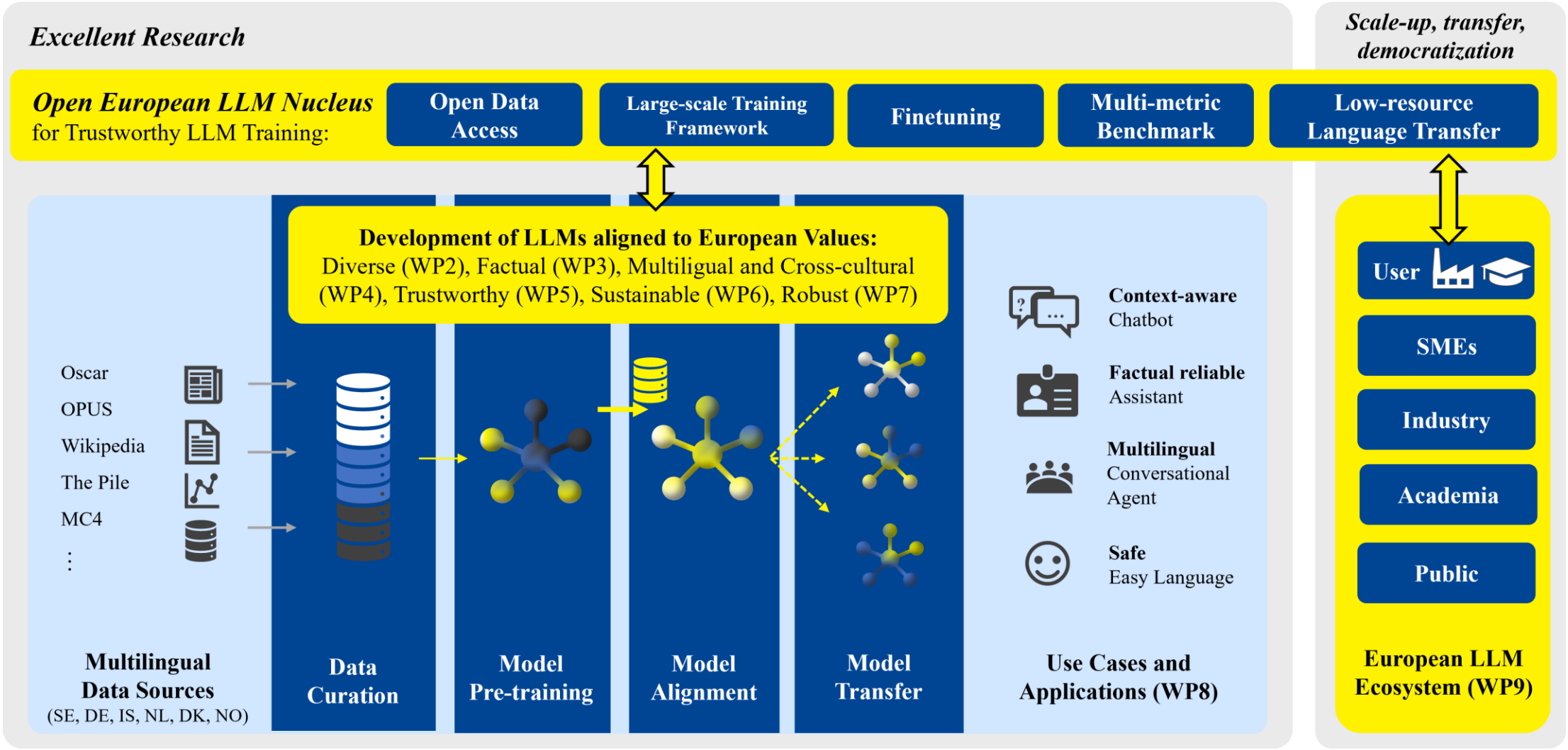
Parameters

Parameters	FLOPs	FLOPs (in *Gopher* unit)	Tokens
400 Million	1.92e+19	1/29,968	8.0 Billion
1 Billion	1.21e+20	1/4,761	20.2 Billion
10 Billion	1.23e+22	1/46	205.1 Billion
67 Billion	5.76e+23	1	1.5 Trillion
175 Billion	3.85e+24	6.7	3.7 Trillion
280 Billion	9.90e+24	17.2	5.9 Trillion
520 Billion	3.43e+25	59.5	11.0 Trillion
1 Trillion	1.27e+26	221.3	21.2 Trillion
10 Trillion	1.30e+28	22515.9	216.2 Trillion

Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion

https://www.lesswrong.com/posts/midXmMb2Xg37F2Kgn/new-scaling-laws-for-large-language-models

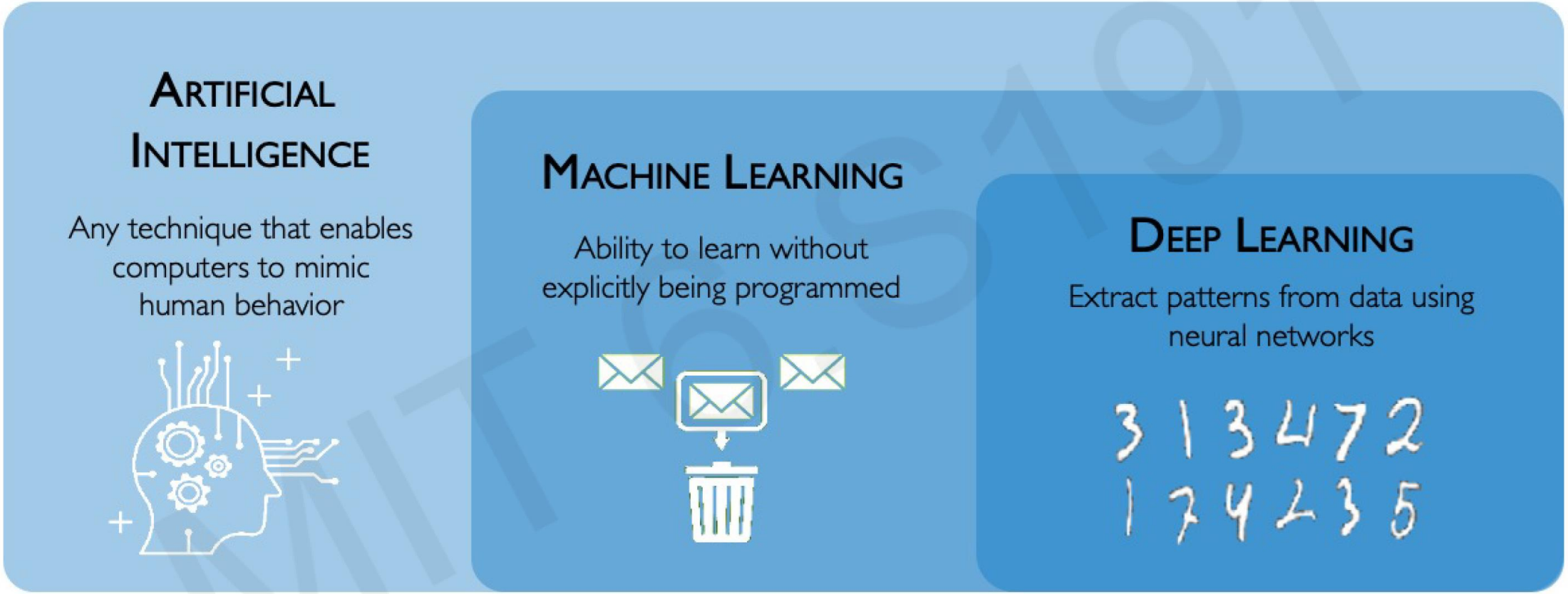
TrustLLM: Project Concept



Large Language Models – The Next Step

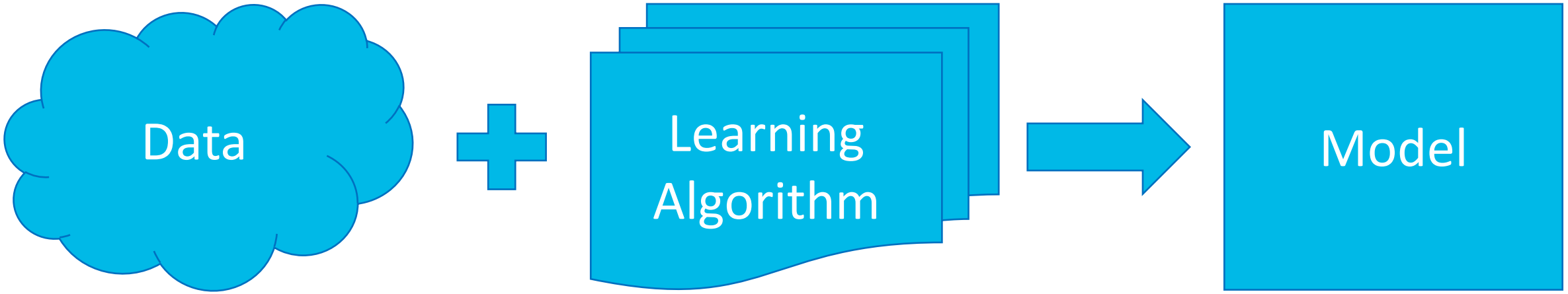
- Making LLMs more factual, coherent, and trustworthy.
- Reduce the need for data and compute, both for training and inference.
- More modular and specialised LLMs for different domains and usages.
- Multi-modal LLMs.
- LLMs can act as a new interface to large amounts of information.
- There is basically infinitely many possible applications thus I expect LLMs will play a major role in most future applications.

What is Deep Learning?

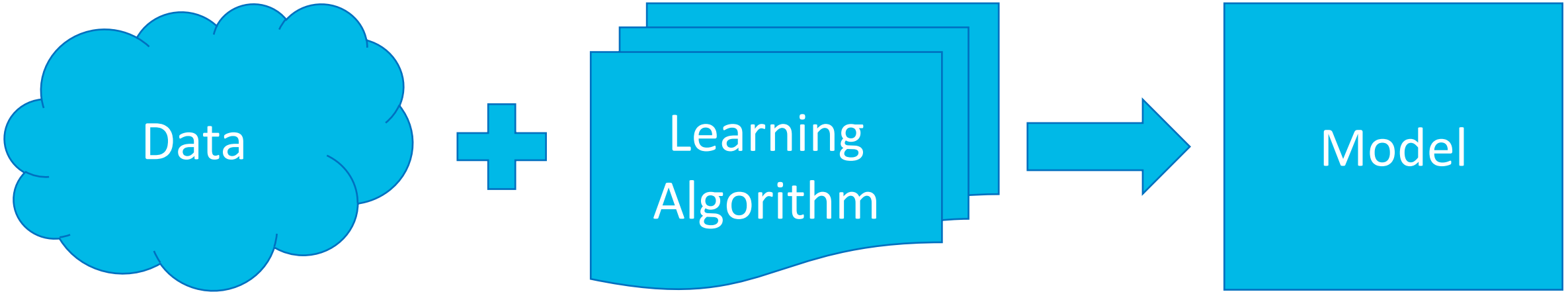


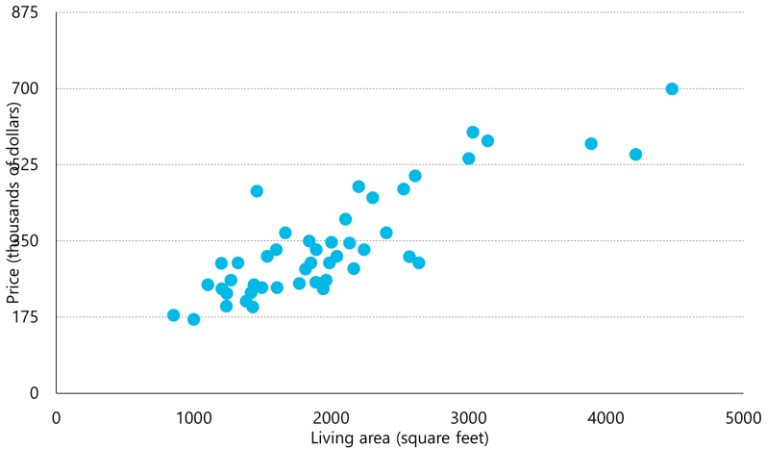
Teaching computers how to **learn a task** directly from **raw data**

Maskininlärning



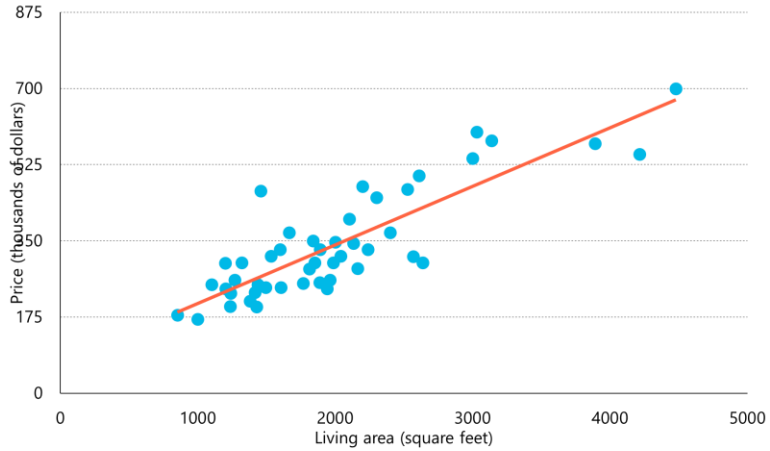
Maskininlärning



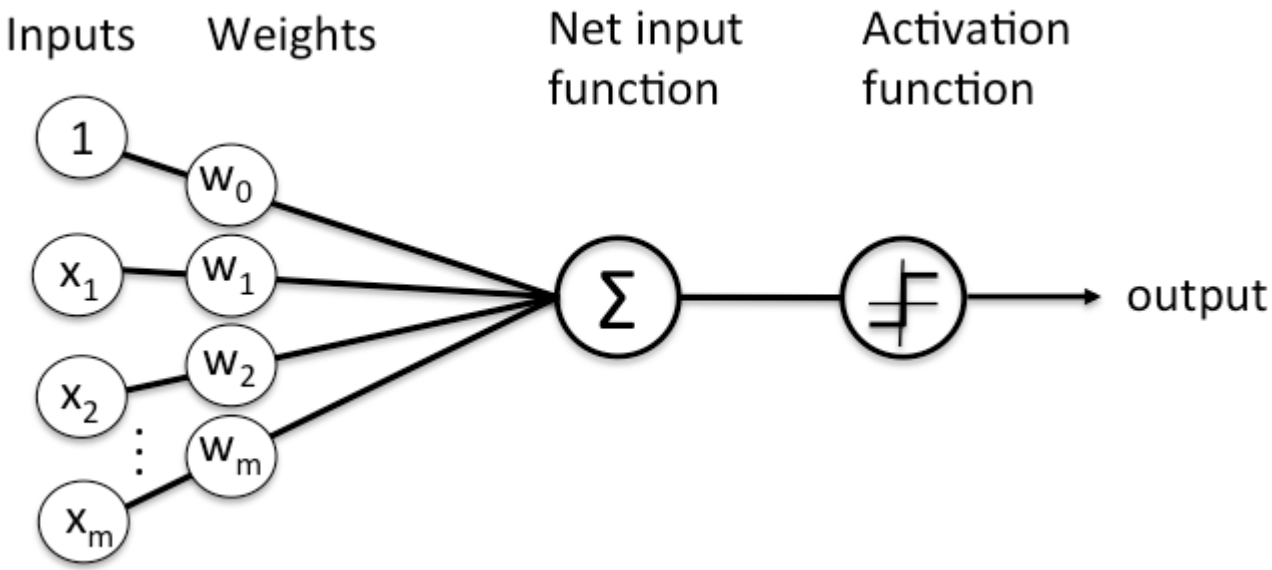


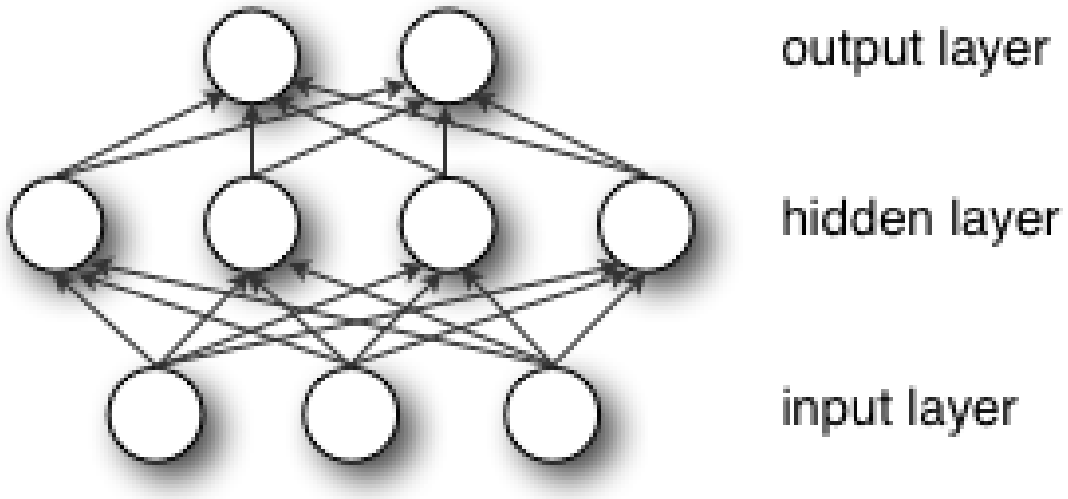
Linjär regression

$$\hat{y} = \boldsymbol{x\theta}$$

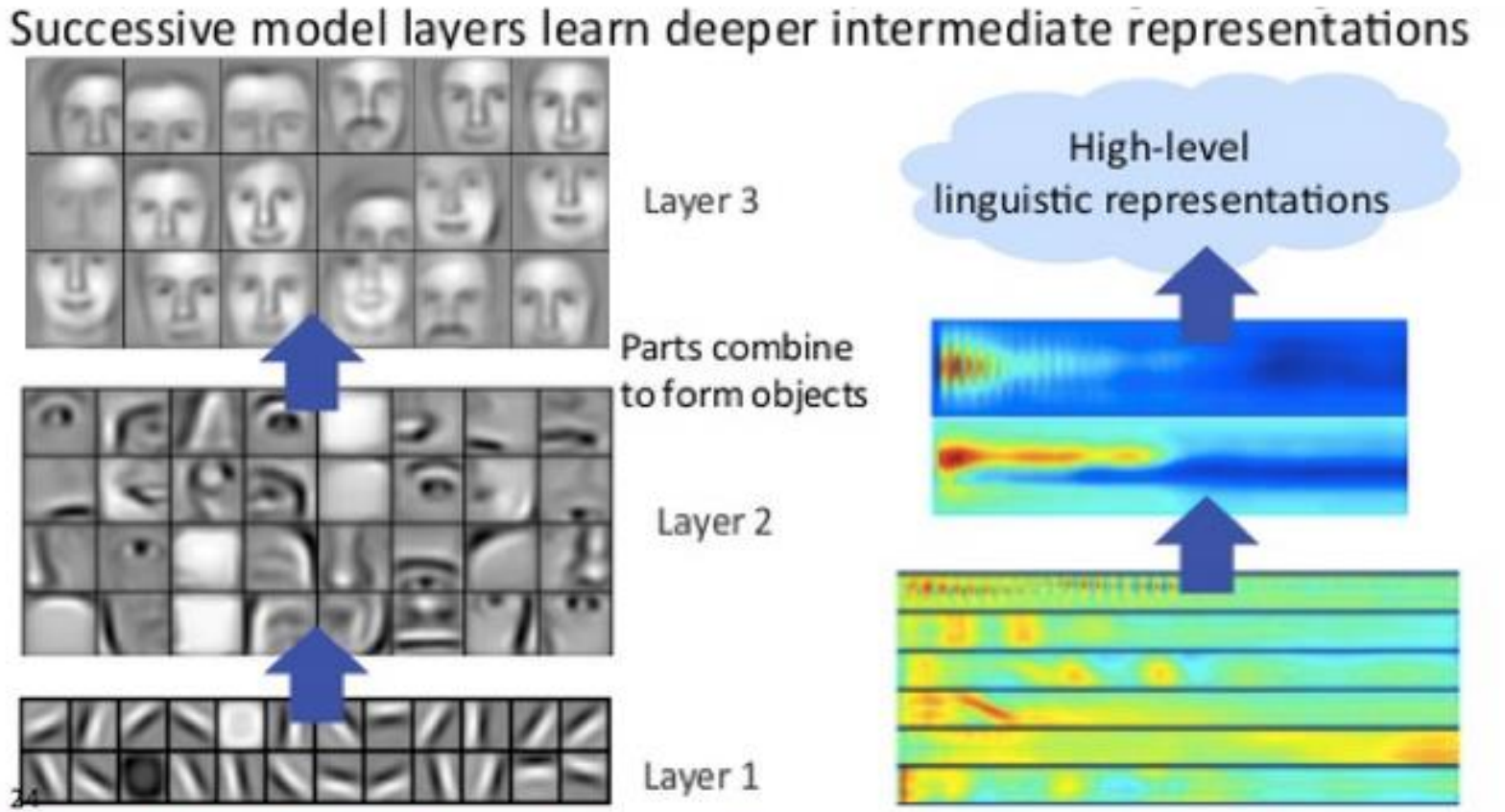


Neural Networks



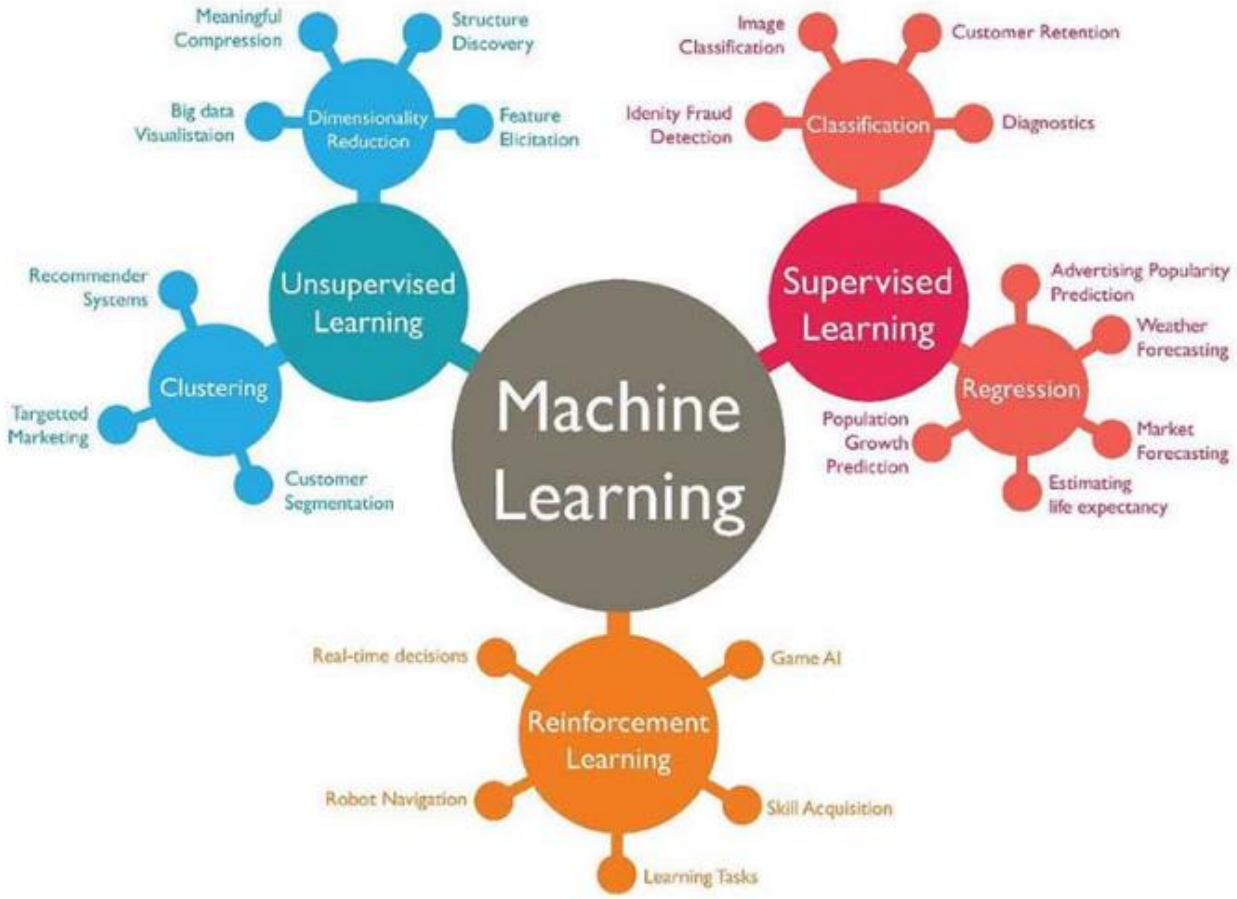


Deep Neural Networks



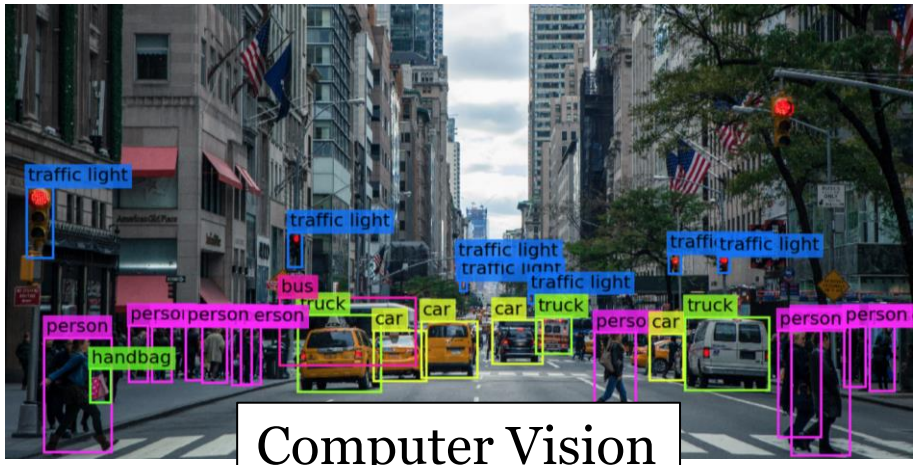
Types of Machine Learning

- Supervised learning
  - Given input-output examples f(X)=Y, learn the function f().
- Unsupervised learning
  - Given input examples, find patterns such as clusters
- Reinforcement learning
  - Select and execute an action, get feedback, update policy (what action to do in which state).



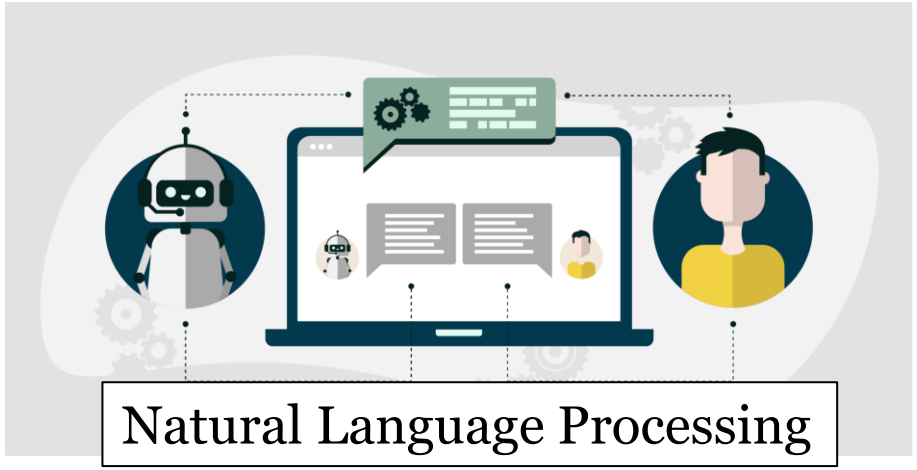
https://www.techleer.com/articles/203-machine-learning-algorithm-backbone-of-emerging-technologies/

Applications of AI

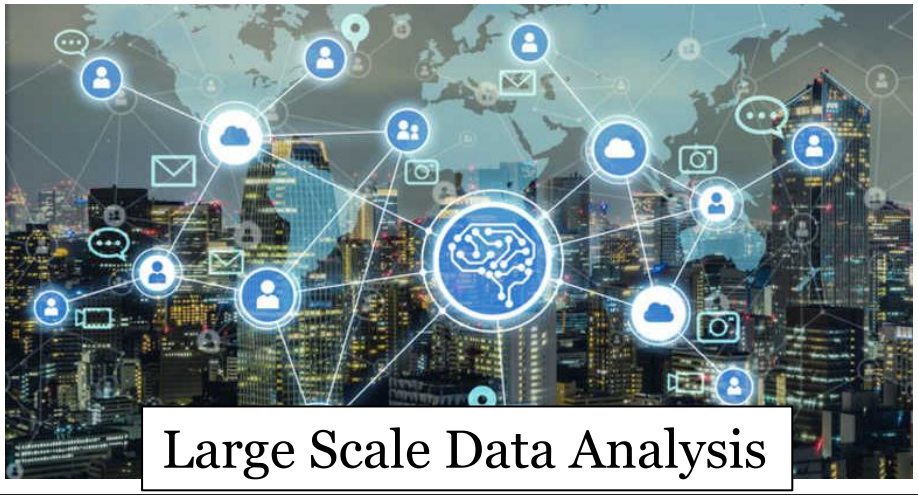


Computer Vision

Voice Recognition



Natural Language Processing

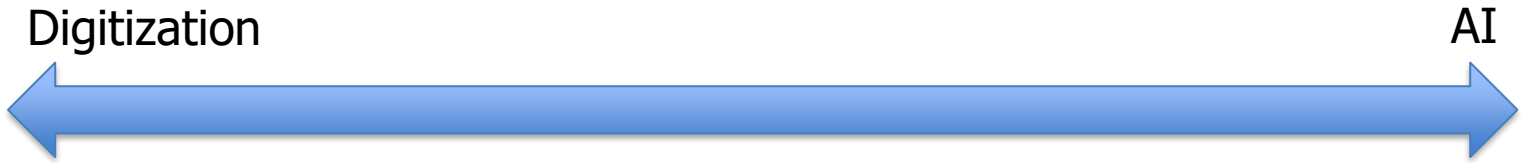


Large Scale Data Analysis

AI and Digitalization

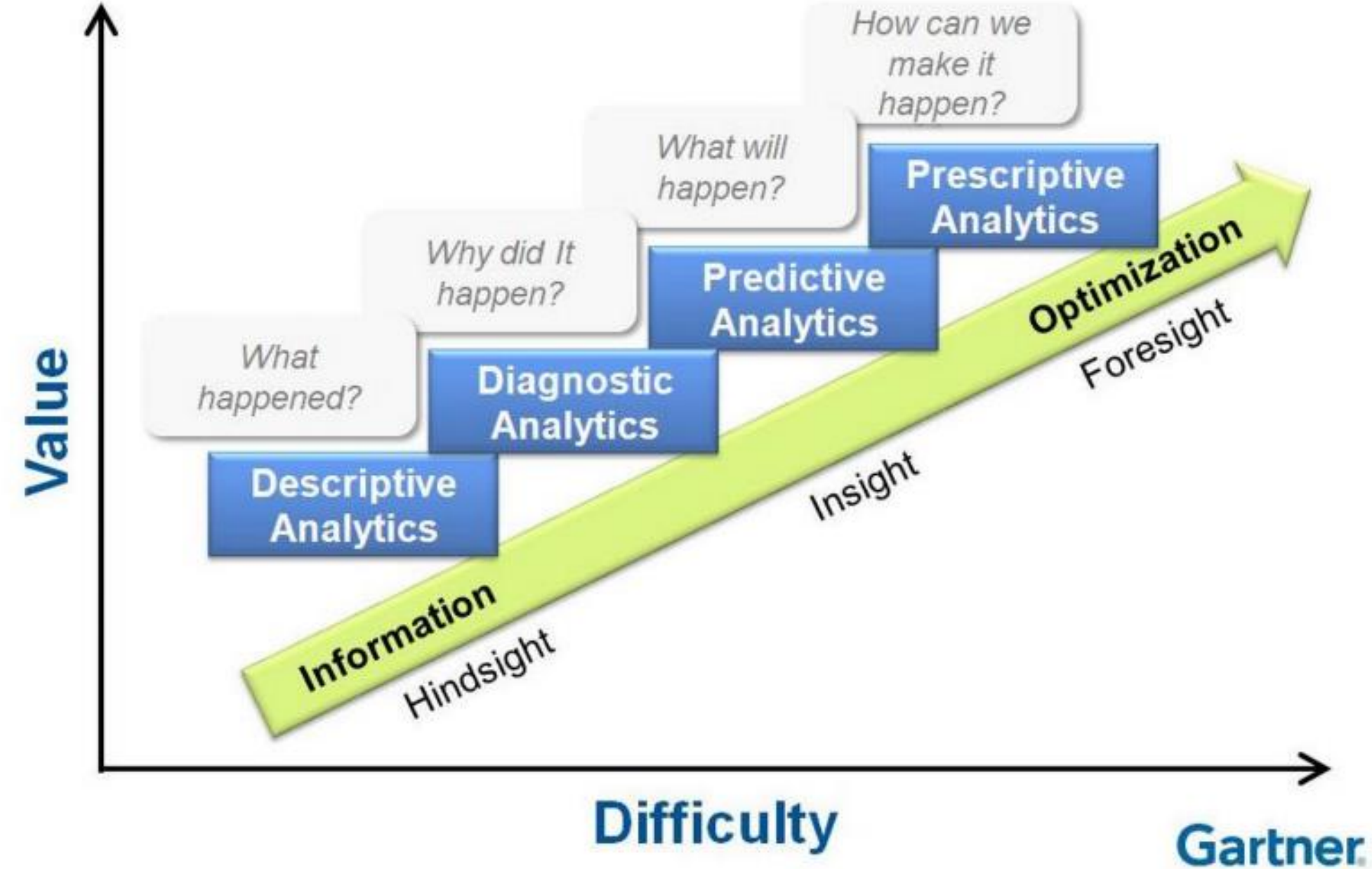
Digitization → **Big Data** → **AI**

first wave | second wave | third wave

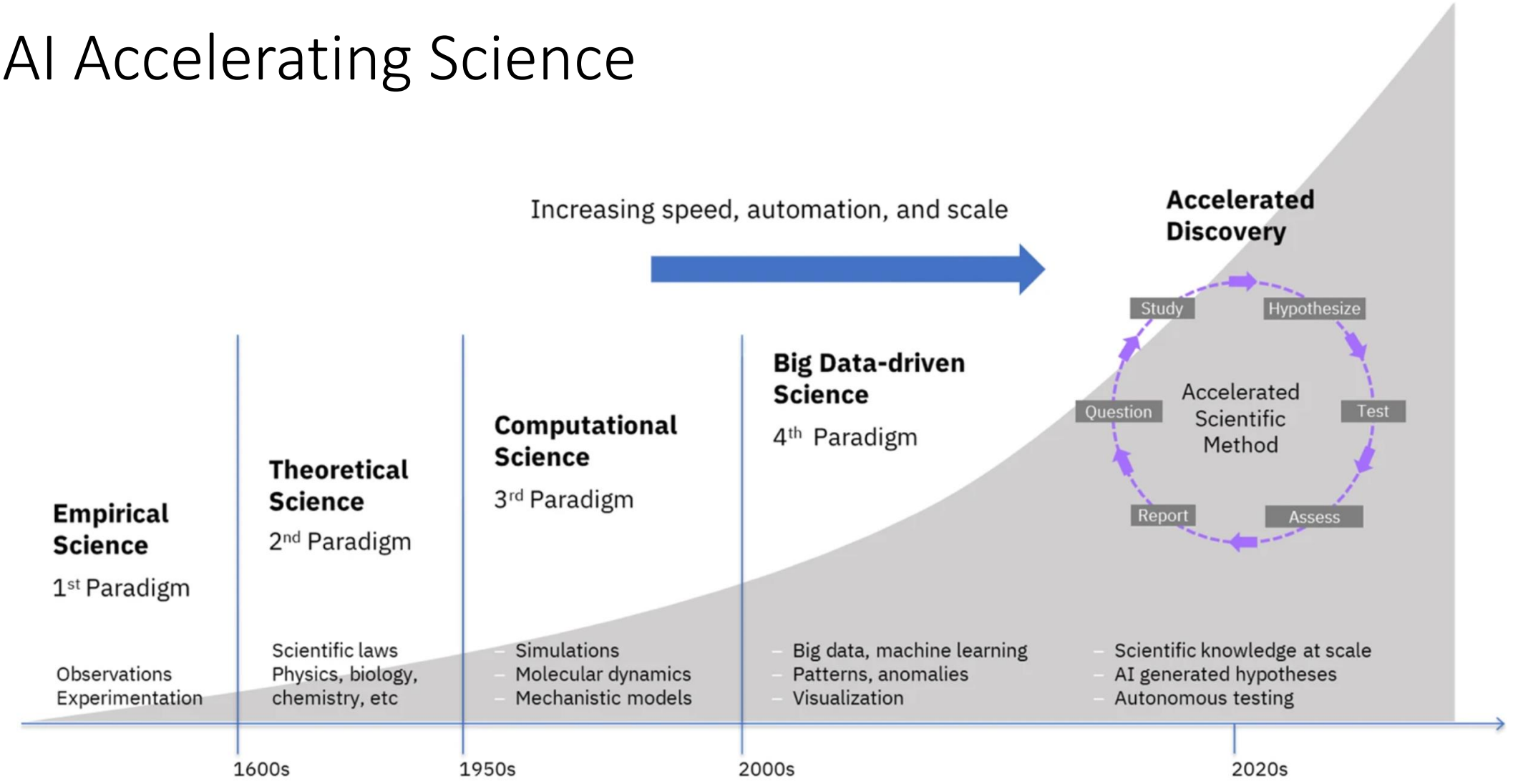


Well defined problems
Predictable situations
Structured data
General solutions
Rationalizes
Evolutionary
...

Hard to define problems
Unanticipated situations
Unstructured data
Adaptable solutions
Amplifies
Revolutionary
...


How can we make it happen?
What will happen?
Why did It happen?
What happened?
Prescriptive Analytics
Predictive Analytics
Diagnostic Analytics
Descriptive Analytics
Optimization
Foresight
Insight
Information
Hindsight
Value
Difficulty
Gartner

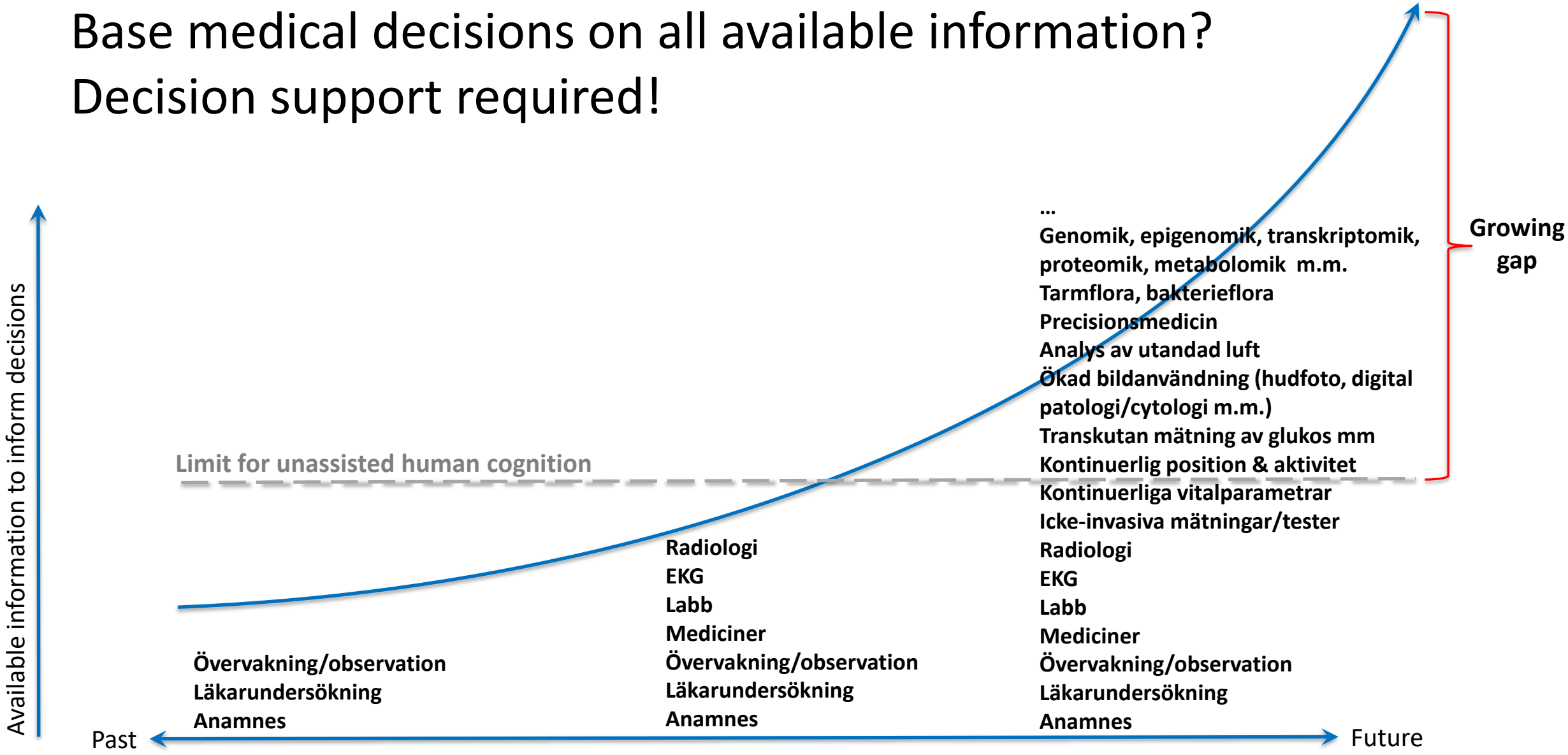
AI Accelerating Science



Science has seen a number of major paradigm shifts, which have been driven by the advent and advancement of core underlying technology.

Pyzer-Knapp, E.O., Pitera, J.W., Staar, P.W.J. *et al.* Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Comput Mater* **8**, 84 (2022). https://doi.org/10.1038/s41524-022-00765-z

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

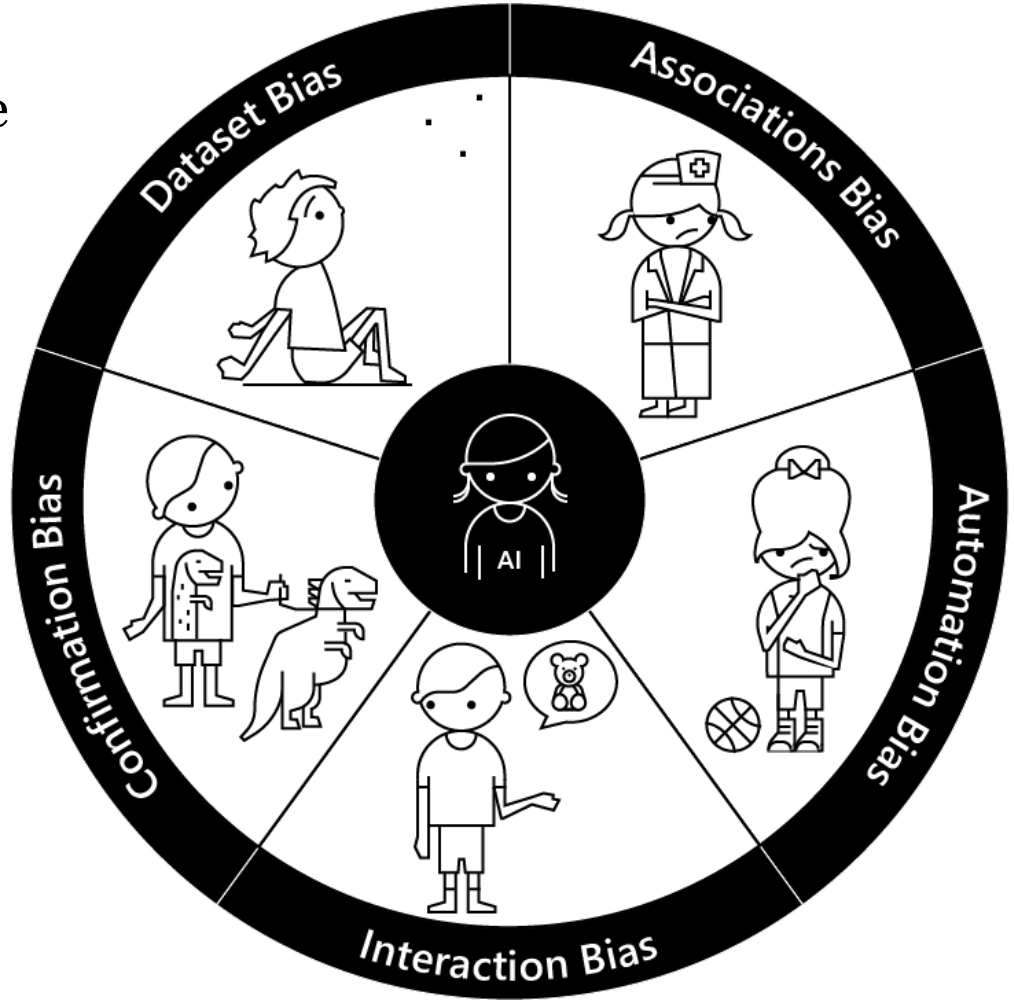


LINKÖPINGS UNIVERSITET

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

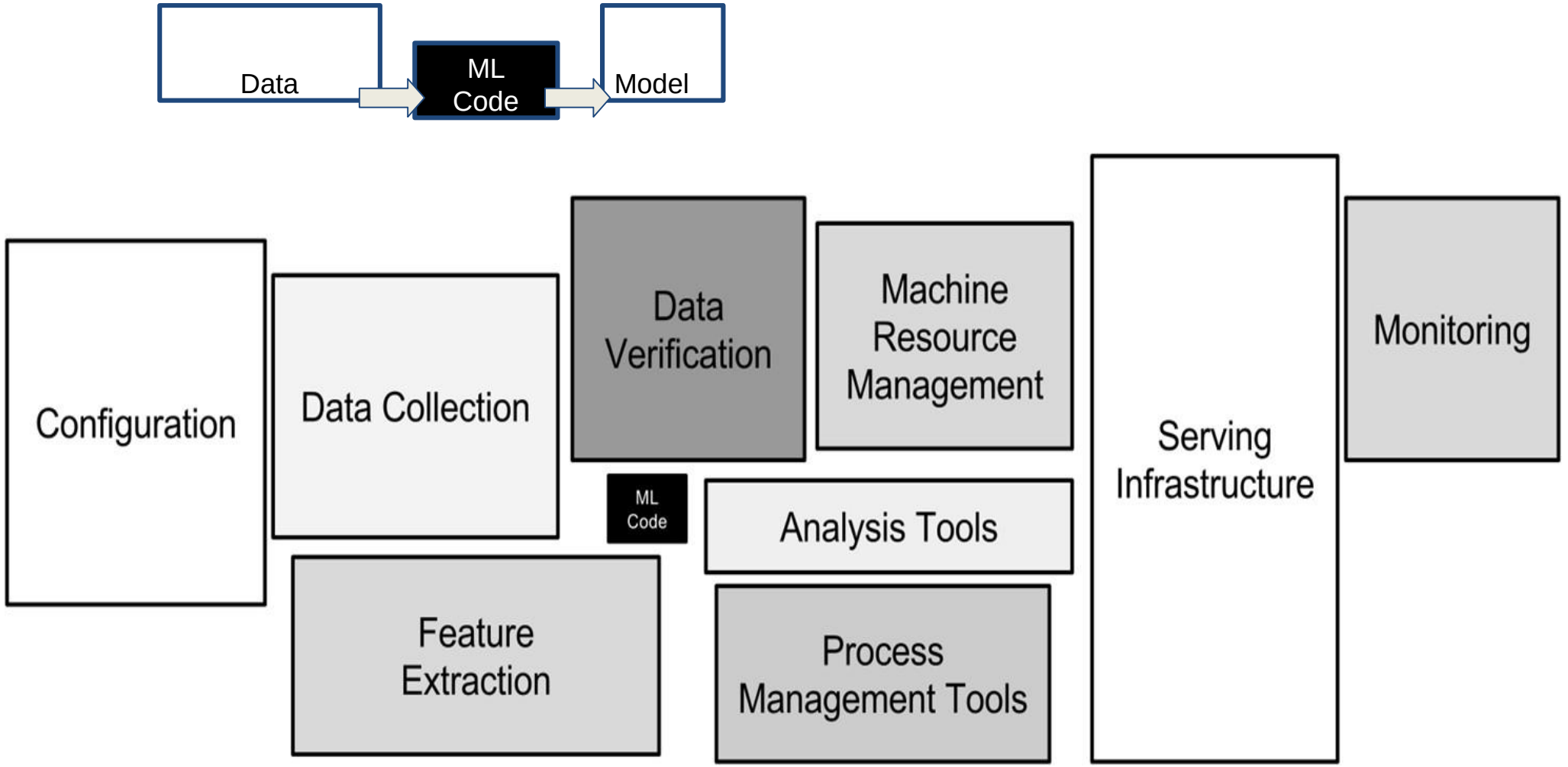
Bias

- **Dataset bias** – When the data used to train machine learning models doesn't represent the diversity of the customer base.
- **Association bias** – When the data used to train a model reinforces and multiplies a cultural bias.
- **Automation bias** – When automated decisions override social and cultural considerations.
- **Interaction bias** – When humans tamper with AI and create biased results.
- **Confirmation bias** – When oversimplified personalization makes biased assumptions for a group or an individual.



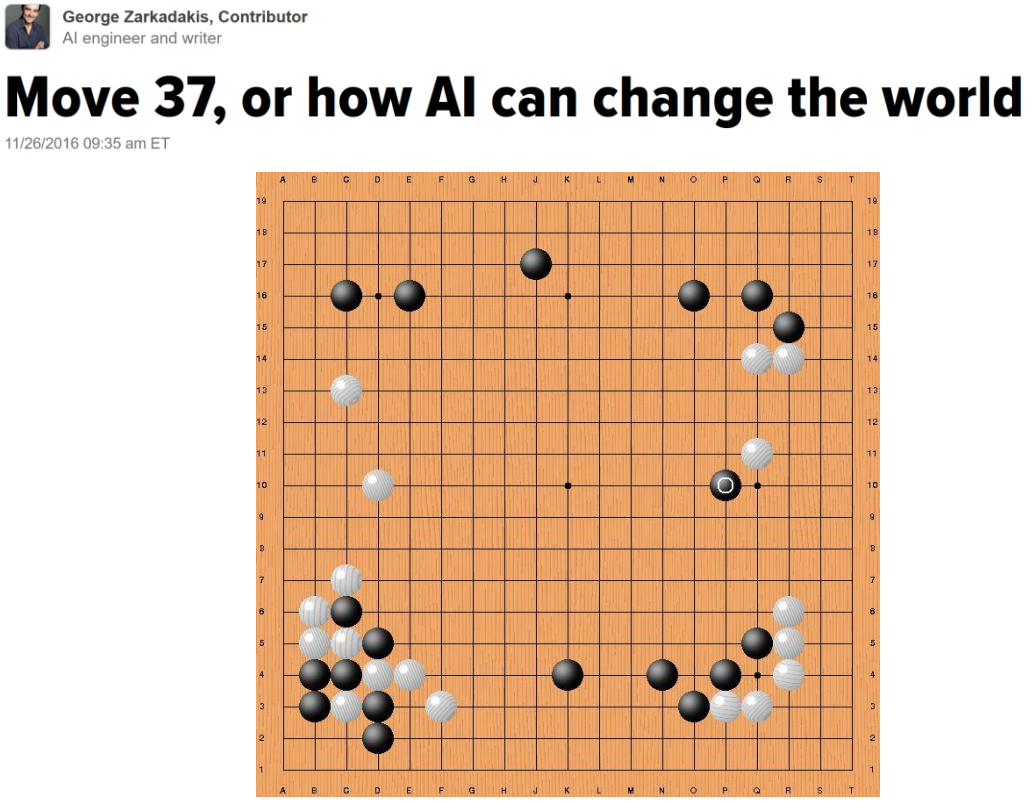
https://medium.com/microsoft-design/how-to-recognize-exclusion-in-ai-ec2d6d89f850

The bigger system / picture



Hidden technical debt in Machine Learning Systems, Sculley et. al. (NIPS 2015)

How to Evaluate AI Systems?



https://www.huffpost.com/entry/move-37-or-how-ai-can-change-the-world_b_58399703e4b0a79f7433b675

Ethics Guidelines for Trustworthy AI – Overview

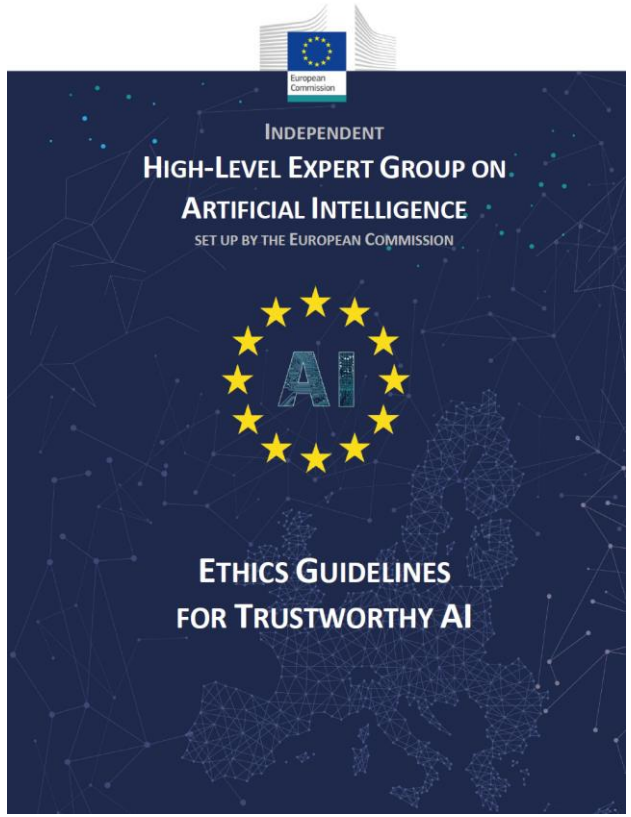
Human-centric approach: AI as a means, not an end

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

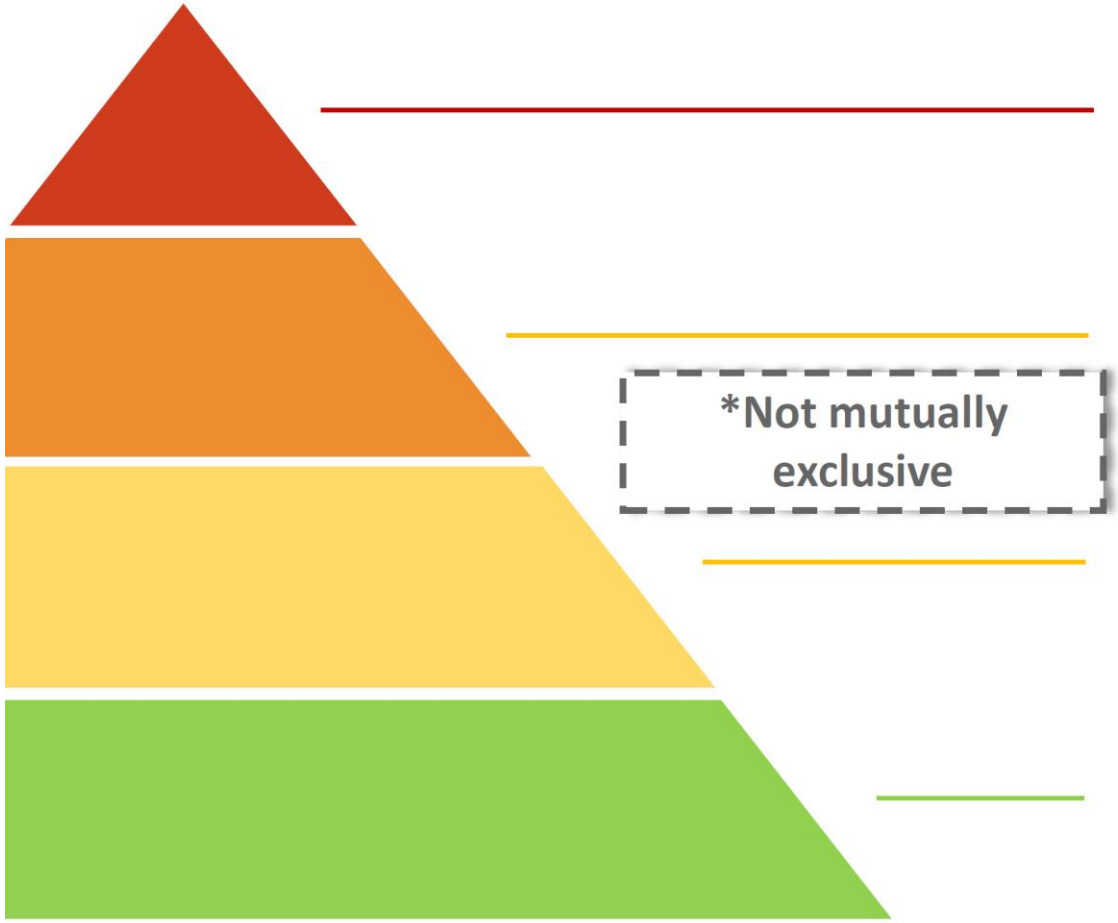
Equal and just distribution of benefits and costs.

Explicability

Transparent, open with capabilities and purposes, explanations

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

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**

AI Act requirements (articles 9-15) Support to building trust in AI as technology TBD 2023, in force 2026	**CEN and CENELEC draft request December 2022 European standards / standardisation deliverables to be drafted in support of trustworthy AI Deadline Jan 2025**	**ISO/IEC standards and/or ongoing standardisation work**
Risk management system (art 9)	Risk management system for AI systems	23894:2023 AI risk management TR 24028:2020 AI trustworthiness *TR 22100-5:2021 Machine safety and AI*
Data quality (art 10)	Governance and quality of datasets used to build AI systems	38505-1:2017 Data governance TR 24027:2021 Bias in AI
Record keeping (art 12)	Record keeping through logging capabilities by AI systems	
Transparency and information to users (art 13)	Transparency and information provisions to the users of AI systems	TR 24028:2020 AI trustworthiness TR 9241-810:2020 Human-system interaction
Human oversight (art 14)	Human oversight of AI systems	TR 24028:2020 AI trustworthiness TR 9241-810:2020 Human-system interaction
Accuracy (art 15)	Accuracy specifications for AI systems	TR 24028:2020 AI trustworthiness
Robustness (art 15)	Robustness specifications for AI systems	TR 24029-1:2021 Robustness of neural networks
Cybersecurity (art 15)	Cybersecurity specifications for AI systems	TR 24028:2020 AI trustworthiness 27001:2022 / 27002:2022 Organisational cybersecurity
Quality management system (art 17)	Quality management system for providers of AI systems, including post-market monitoring process	38507:2022 Organisational governance of AI
Conformity assessment (art 43)	conformity assessment for AI systems	

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

TAILOR – Vision

Develop the scientific foundations for Trustworthy AI

integrating learning, optimisation and reasoning

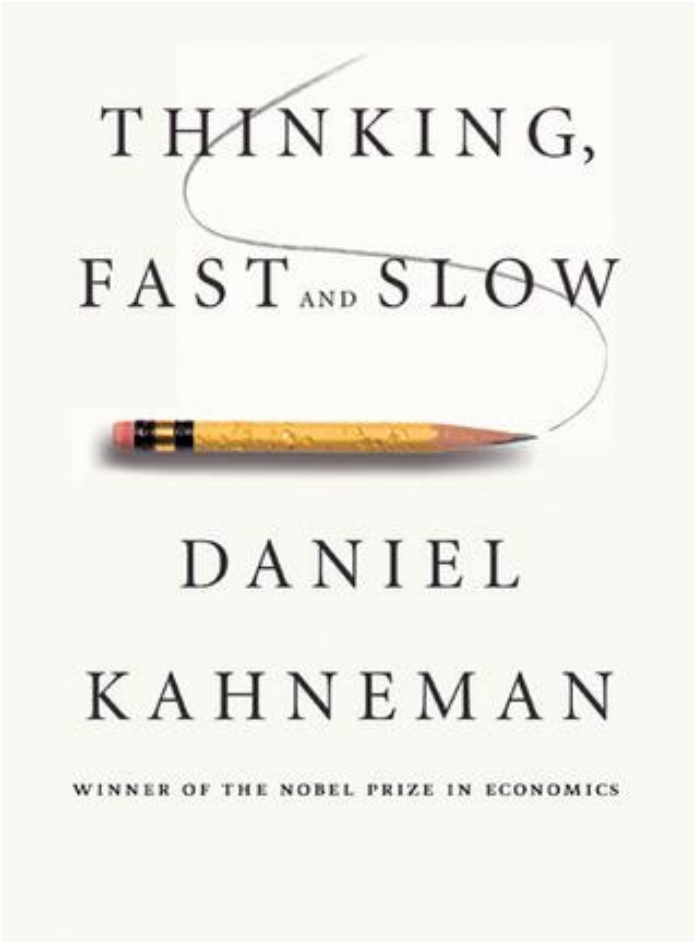
to realise the European vision of human-centered Trustworthy AI.

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

Boosting Capacity to Tackle Major Scientific Challenges

- 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

STRATEGIC RESEARCH & INNOVATION ROADMAP OF TRUSTWORTHY AI

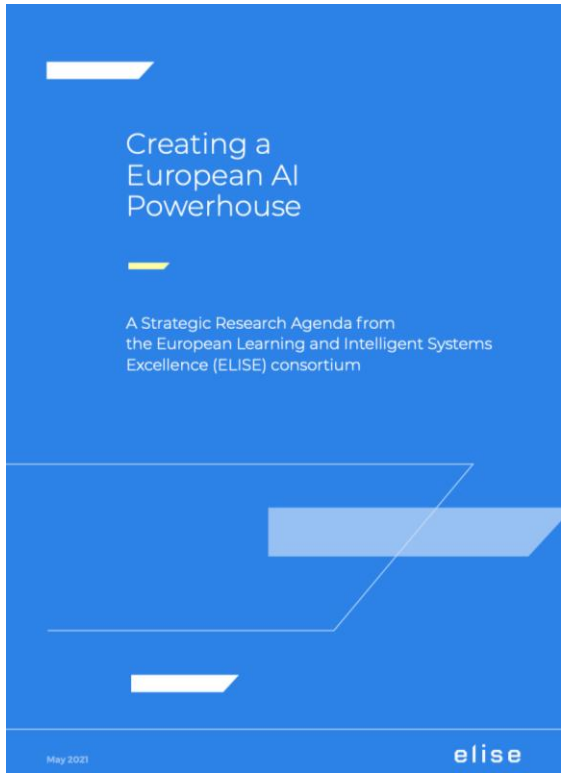
The Scientific Foundations of Trustworthy AI in Europe for the Years 2022-2030

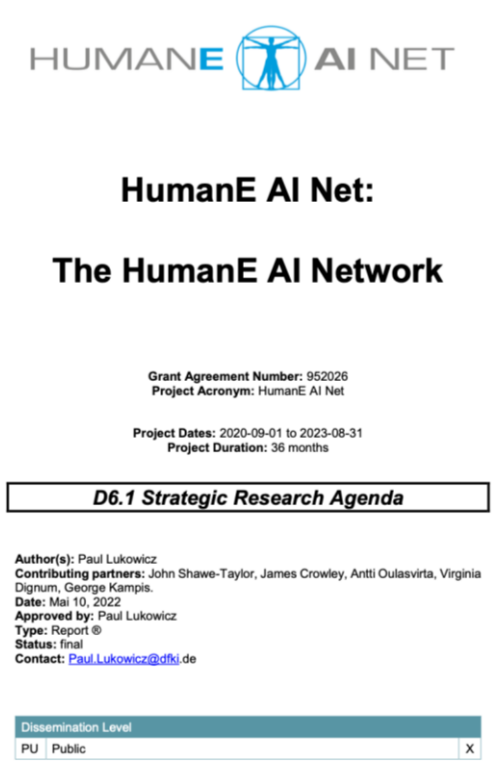
TAILOR Strategic Research and Innovation Roadmap (SRIR) aims to boost research on Trustworthy AI by clearly defining the major research challenges.

https://tailor-network.eu/research-overview/strategic-research-and-innovation-roadmap/

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

Joint Strategic Research Agenda





Joint Editorial Board: AI4Media; ELISE; ELSA; euROBIN; HumaneAI; TAILOR; VISION

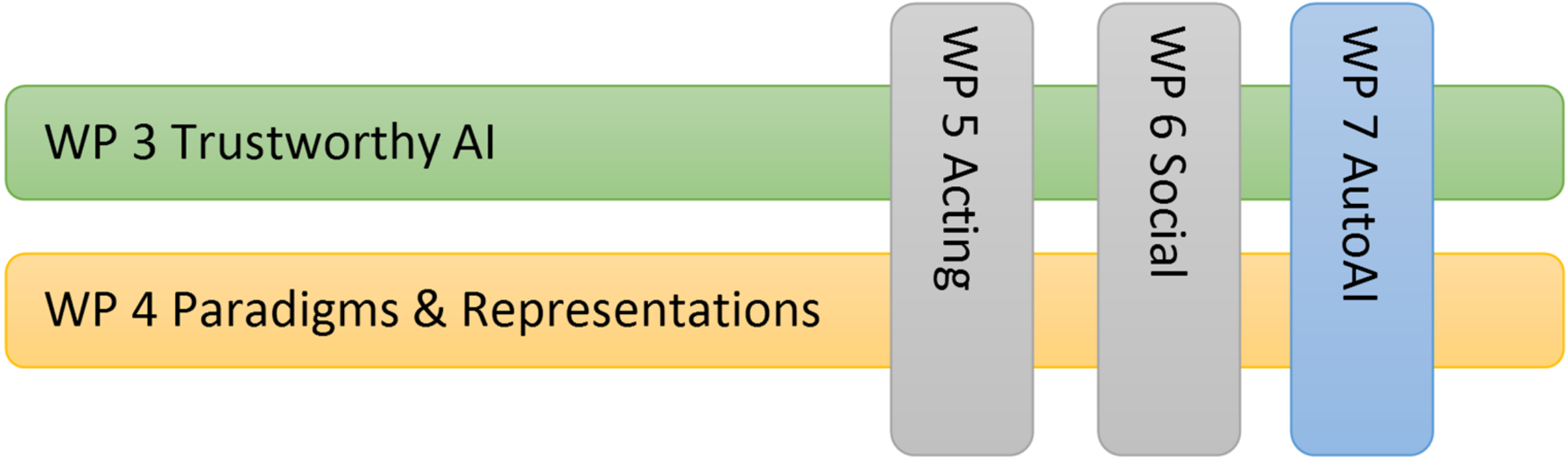
Joint SRA – Research Challenges

1. Building the technical foundations of safe and trustworthy ADR
2. Integrating AI into deployed or embedded systems, including robots
3. Enhancing human capabilities with collaborative ADR
4. Accelerating research and innovation with ADR
5. Understanding interactions between ADR, social needs and socio-technical systems
6. Advancing fundamental theories, models, and methods
7. Ensuring legal compliance of ADR systems
8. Advancing hardware for safe and energy efficient interaction between ADR technologies, humans, and the environment

Joint SRA – Research Topics Covered

Human agency and oversight
Robustness
Safety
Privacy and data governance
Transparency, explainability, and human-understandable AI
Diversity, non-discrimination and fairness
Societal and environmental wellbeing
Accountability
Trade-offs and interactions
AutoML and AutoAI
Verification, validation, and certification
Traceability
AI at the edge
Human-centric AI
Entanglement between AI, software and hardware
Multi-agent collaborations
Models of human-AI collaboration
Common ground and shared representations
Active learning, lifelong learning, and dynamic feedback
Knowledge representations
Understanding intentions
Generative AI
Human-AI co-evolution
Simulation and emulation
Causal AI
Encoding domain knowledge
Multimodal learning
AI for research and innovation
Research methodology and infrastructure
Data and robotics stewardship
Participatory design
Responsible research and innovation
Dynamics of socio-technical systems
AI impact on fundamental rights and society
X – by – design
Foundation models
Learning strategies: active learning, deep learning, reinforcement learning, transfer learning, few-shot learning, federated learning, continual learning, multimodal learning, causal inference
Computer vision
Natural Language Processing
Quantum computing and machine learning
Integration of learning methods
AI for regulation
Next generation soft robotics
Next generation electronics for physical interaction
Next generation actuation technologies
Physics-enabled digital twins

TAILOR – Basic Research Program



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

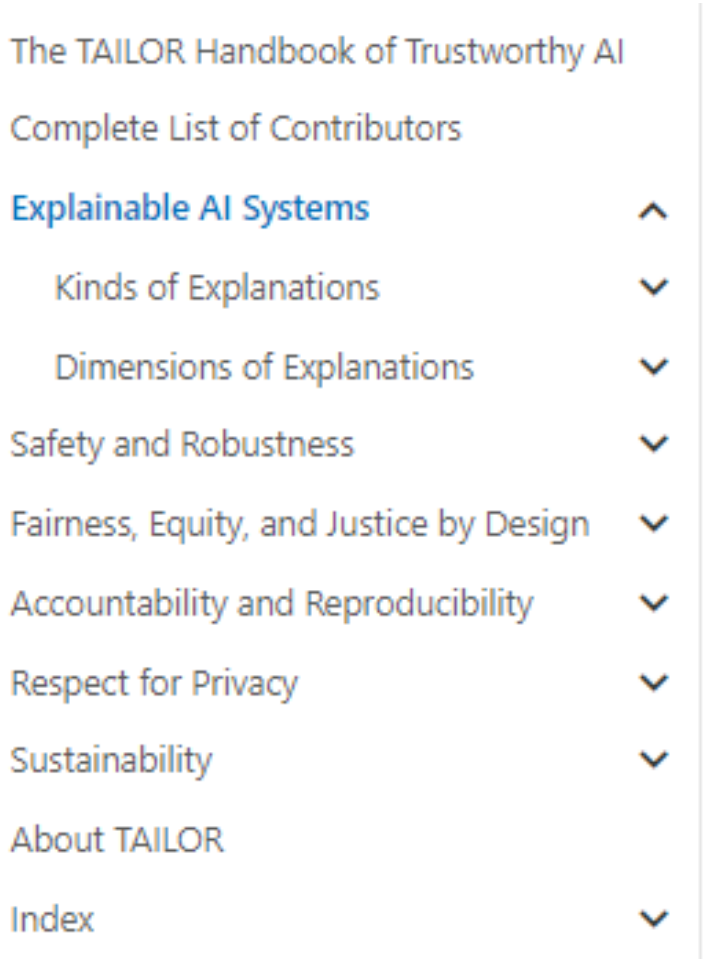
Trustworthy AI – TAILOR Perspective

- 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

Trustworthy AI Handbook

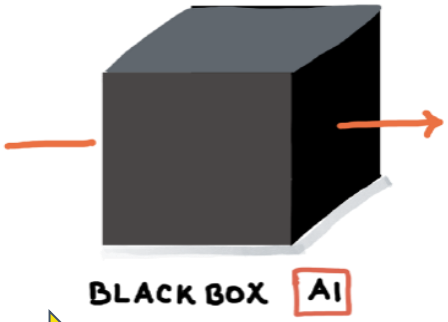
- An **online encyclopedia** of the major scientific and technical terms related to Trustworthy AI
- Contains an overview of the **main dimensions of trustworthiness**, major challenges and solutions in the field, and the latest research developments
- For **non experts, researchers and students**
- 30 contributors from all areas of Trustworthy AI
- Integrated process for enrichment of Wikipedia while maintaining the integrity of the Handbook
- 1st version available: https://tailor-network.eu/handbook/

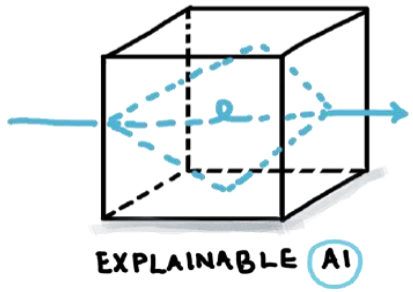


This project is funded by the EC under H2020 ICT-48

WP3: Survey Explainable AI Systems

- Benchmarking and Survey of Explanation Methods for Black Box Models
- Many positive implications:
  - Understand the internal reasoning of the model
  - Identify bias, errors and problems
  - Develop better models





What kind of data has been used as input ?

1. Tabular 2.Images 3.Text

What kind of explanations have been considered ?

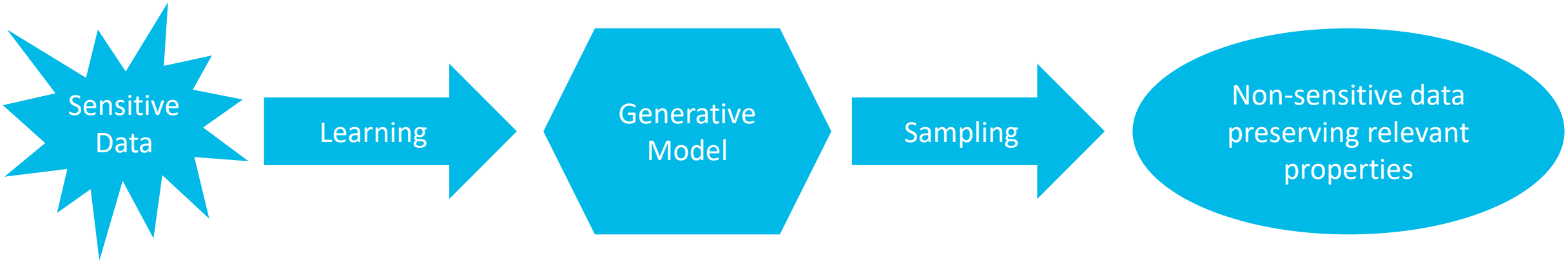
- Tabular
  - Feature importance
  - Rules/Prototype
  - Counterfactuals
- Images
  - Saliency map
  - Prototypes
  - Counterfactuals
- Text
  - Sentence highlighting
  - Attention-based methods

F. Bodria, F.Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, S. Rinzivillo. Benchmarking and Survey of Explanation Methods for Black Box Models. https://arxiv.org/abs/2102.13076

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

Privacy-preserving synthetic data generation

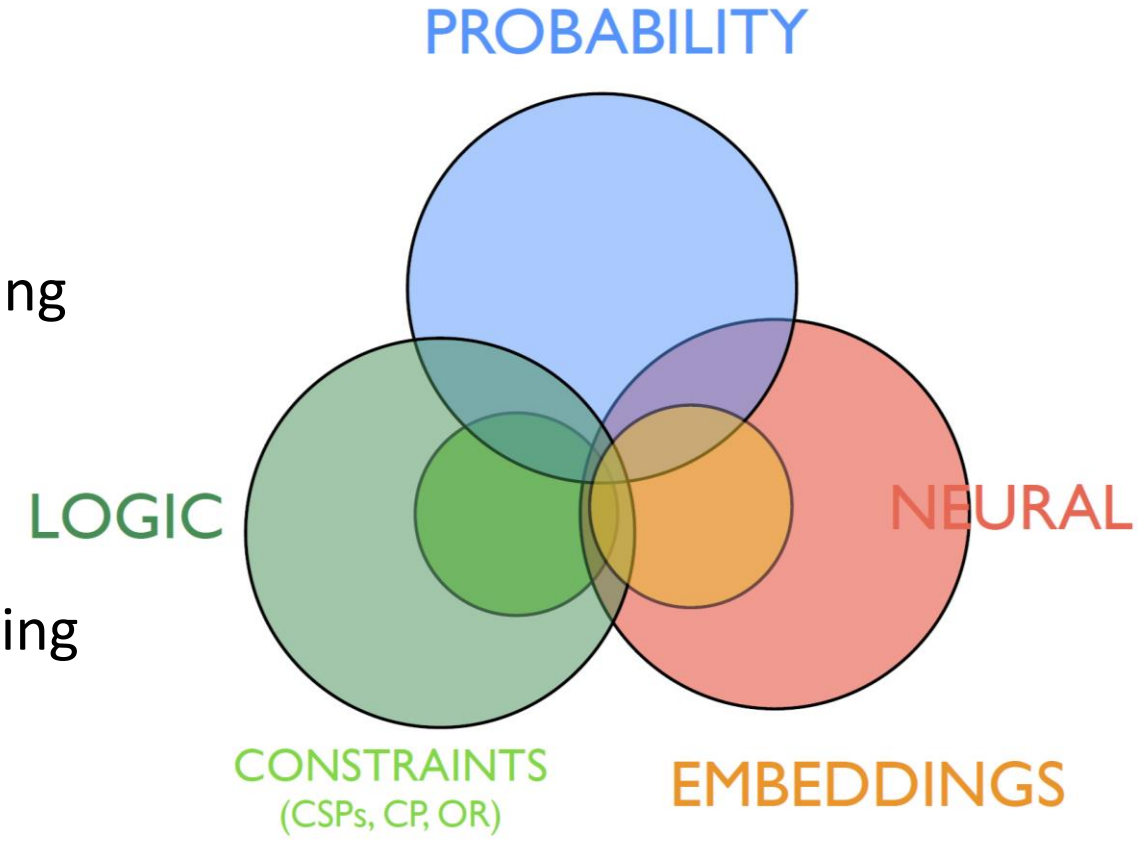
[D. Bergström, Md F. Sikder, R. Ramachandranpillai]



1. Learn a generative model that captures the probability distribution of the sensitive data
2. Create a synthetic data set from the generative model that both captures the salient features of the original data set **and** is non-sensitive
3. Methods for verifying that the synthetic data set is accurate enough
4. Methods for verifying that the synthetic data set is non-sensitive

Paradigms and Representations

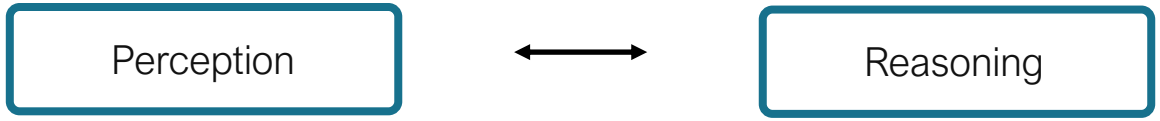
- Goals:
  - Integrate these paradigms
  - Integrate the involved communities
  - Covers five core different communities including
    - Deep & Probabilistic Learning
    - Neuro-Symbolic Computation (NeSy)
    - Statistical Relational AI (StarAI)
    - Constraint Programming & Machine Learning
    -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



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

Neuro-Symbolic Learning

INTEGRATION



INGREDIENTS

COMPLEXITY OF INTEGRATION

7 dimensions for integration

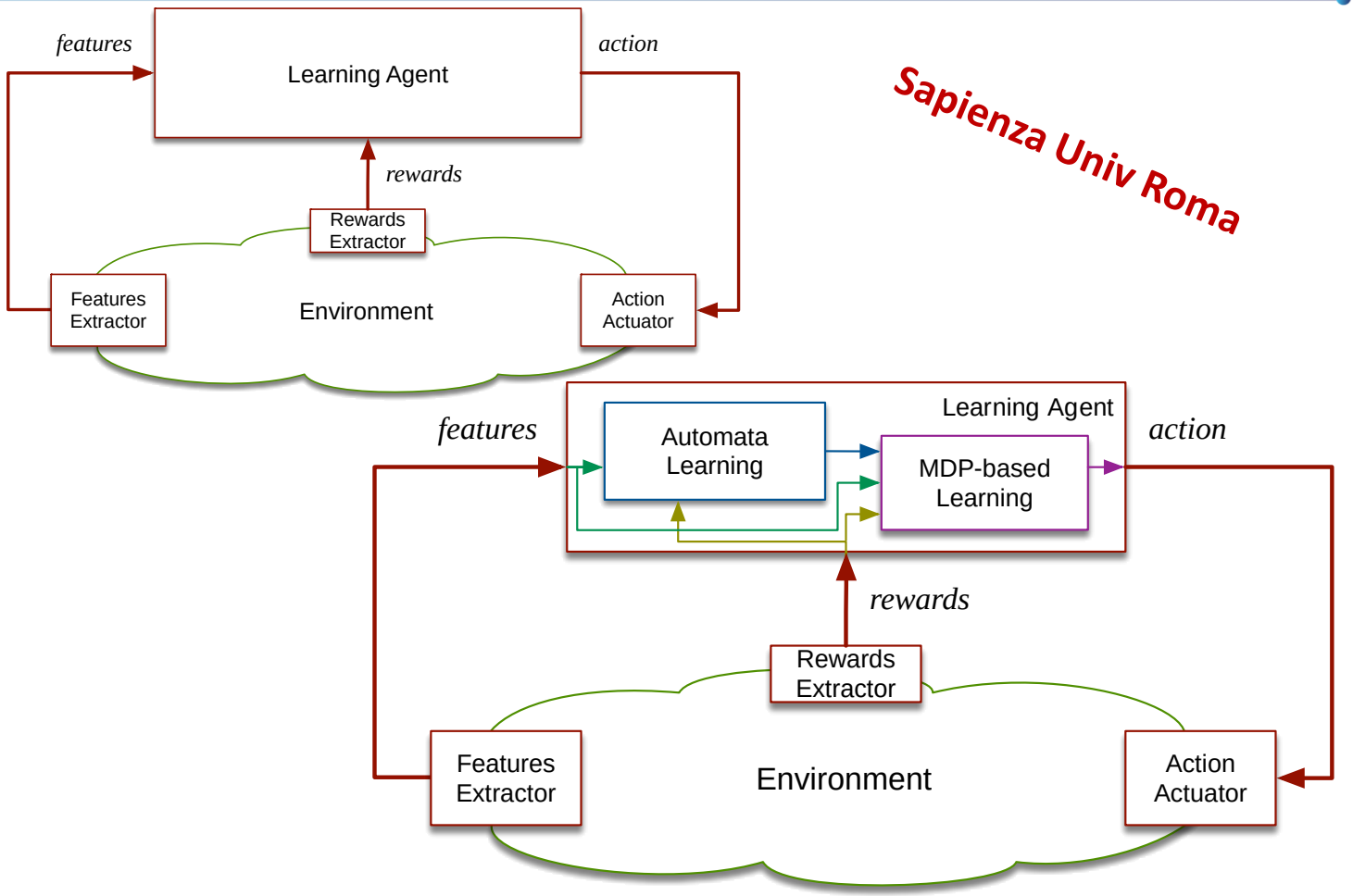
1. Underlying Structure (directed/undirected)
2. Inference (grounding/proof-oriented)
3. Paradigms (Logic, Probability, Neural)
4. Representations (Symbolic, SubSymbolic)
5. Discrete/Continuous (Boolean vs Fuzzy)
6. Learning (Parameter versus Structure)
7. Type of Logic (Propositional, First Order Logic or other forms)

"From Statistical Relational Learning to Neuro-Symbolic Artificial Intelligence" IJCAI 2020

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

WP5 Challenge: Reinforcement Learning in non-Markovian Domains

- Reinforcement Learning in non-Markovian Domains
- Based on Regular Decision Processes (RDP) instead of MDPs
- Handle non-Markovian dynamics (i.e., depending on the history) without postulating a priori existence of hidden variable, as in POMDPs!
- RL on RDPs requires simultaneously learning an automaton for the dynamics and an optimal policy wrt rewards:
  - Polynomial PAC-learnability
  - With no prior knowledge



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

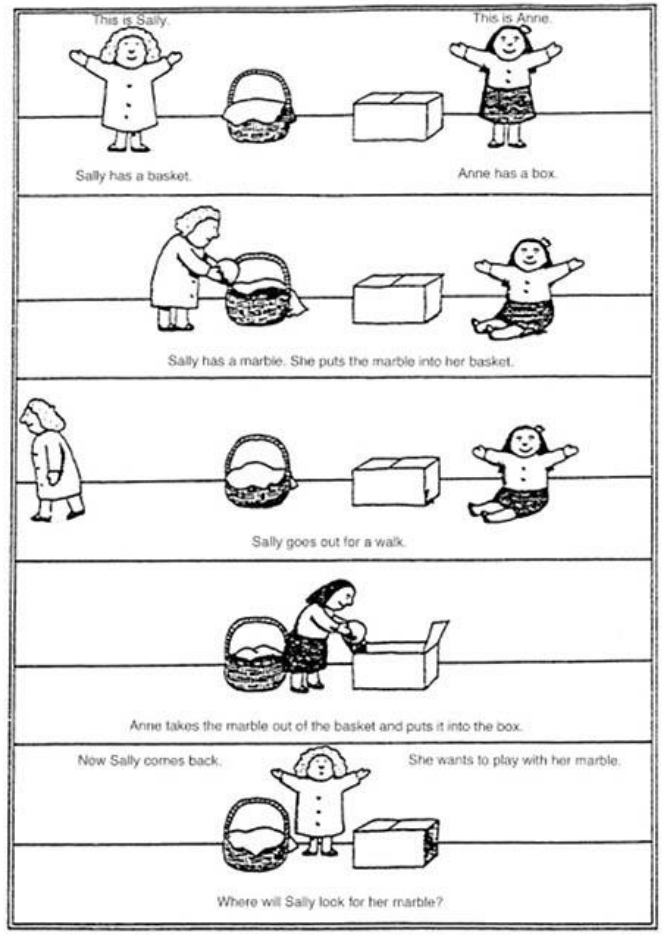
E. Abadi, R. Brafman. Learning and Solving Regular Decision Processes. IJCAI 2020

A. Ronca, G. De Giacomo. Efficient PAC Reinforcement Learning in Regular Decision Processes. IJCAI 2021

Acting meets Social AI

- Topic: acting and planning with other agents' beliefs and goals
  - Theory of Mind = "put oneself in another agent's shoes"
  - Important for any kind of social interaction ➔ related to WP6
    - False-belief tasks (Sally-Ann-Task)
    - Deception, lying,…
    - 'Social intelligence' (otherwise: mind-blind agents)
- Main problem: undecidability of epistemic planning
- Contribution of IRIT: design of a framework for decidable epistemic planning
  - Lightweight fragment of standard epistemic logic
  - Parallel actions
  - Proposal of benchmarks

IRIT, Toulouse



M.C.Cooper, A.Herzig, F.Maffre, F.Maris, E.Perrotin, P.Régnier "A lightweight epistemic logic and its application to planning". Artif. Intell. 2021

M.C.Cooper, A.Herzig, F.Maris, E.Perrotin, J.Vianey "Lightweight Parallel Multi-Agent Epistemic Planning". KR 2020

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

Application: Policy Making and Urban Planning

- Use ABM is to study the effects of policy changes.
- Case: understand policies on the Amsterdam residential dynamics, especially the short-term tourist accommodation market.
- In order to provide insights into qualitative policy effects, we develop a micro-level agent-based simulation. Our spatial model simulates residential migration based on income and house pricing.

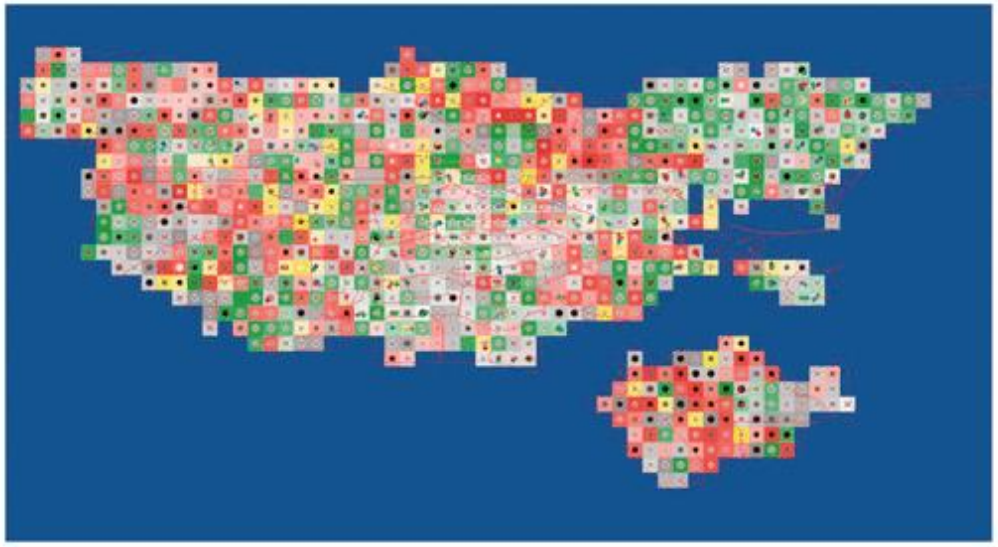
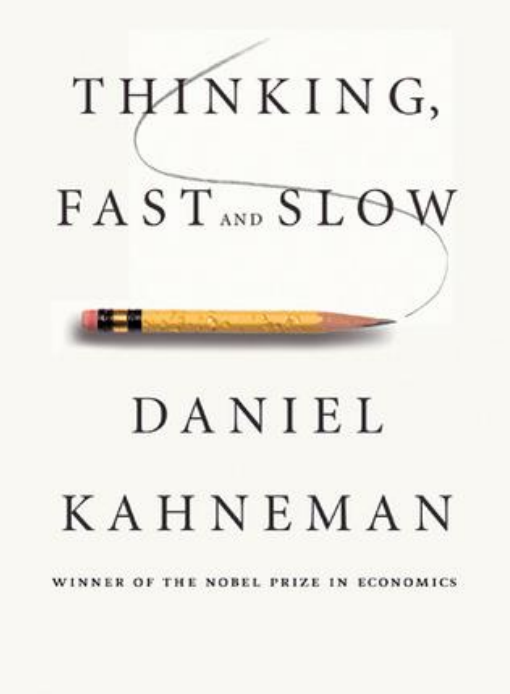


Figure 1. Agent-based model of Amsterdam. Each cell is a residential location. Privately owned locations that are available for touristic rental are green, those that are not available are coloured grey.

Overwater, A., & Yorke-Smith, N. (2021). Agent-based simulation of short-term peer-to-peer rentals: Evidence from the Amsterdam housing market. Environment and Planning B: Urban Analytics and City Science, Sage, March 2021

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

The Way Forward



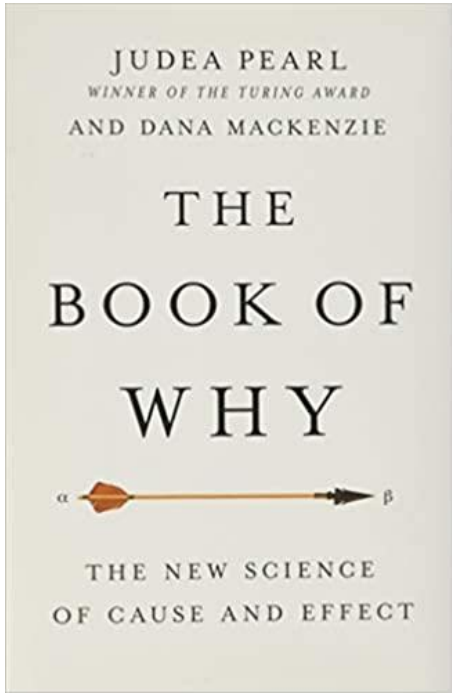
Data

Knowledge/
Assumptions

Causal Models
Combining reasoning
and learning

Explanations

Predictions



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

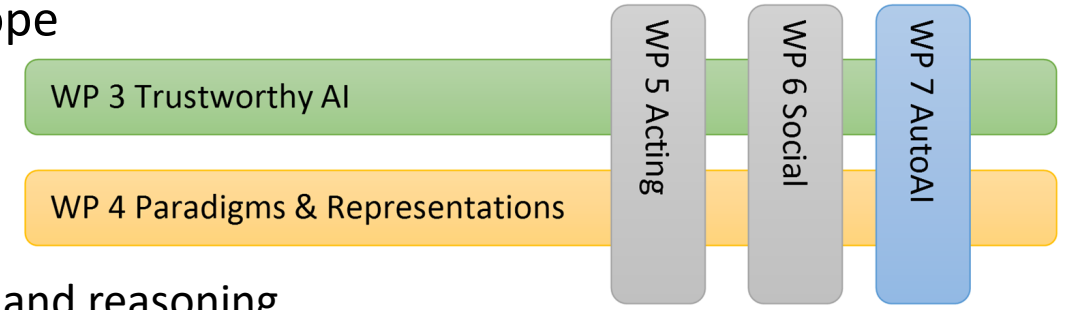
TAILOR ICT-48 Network

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

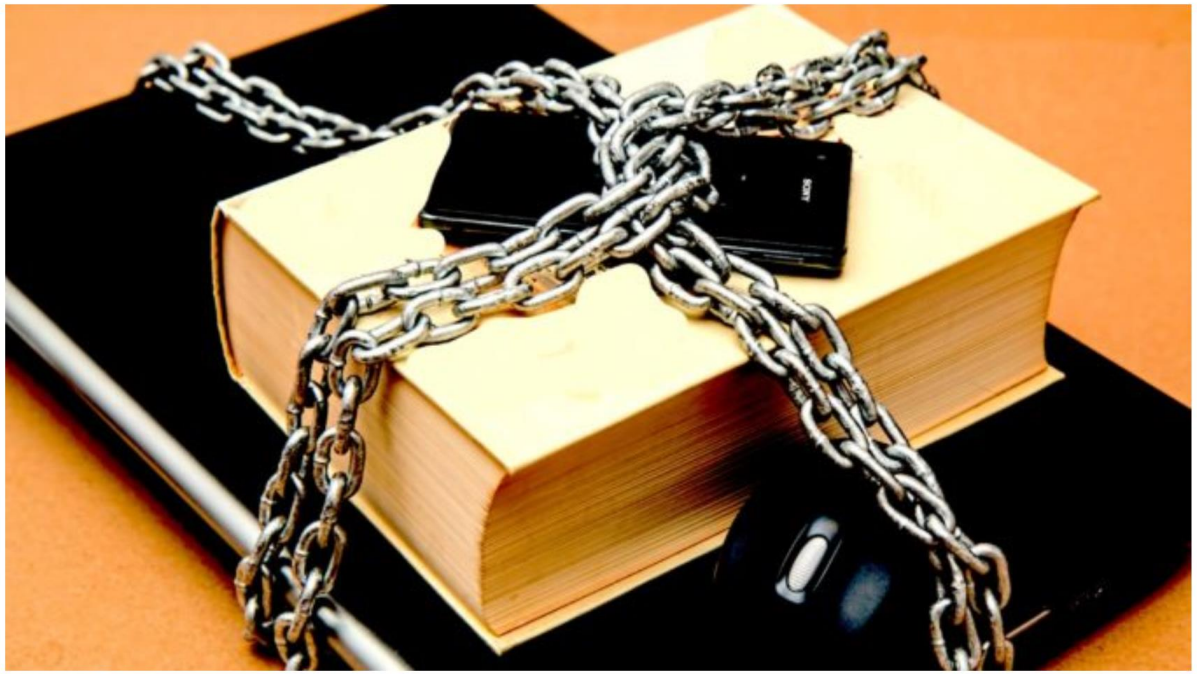


This project is funded by the EC under H2020 ICT-48

External Analysis of Human Decision Making

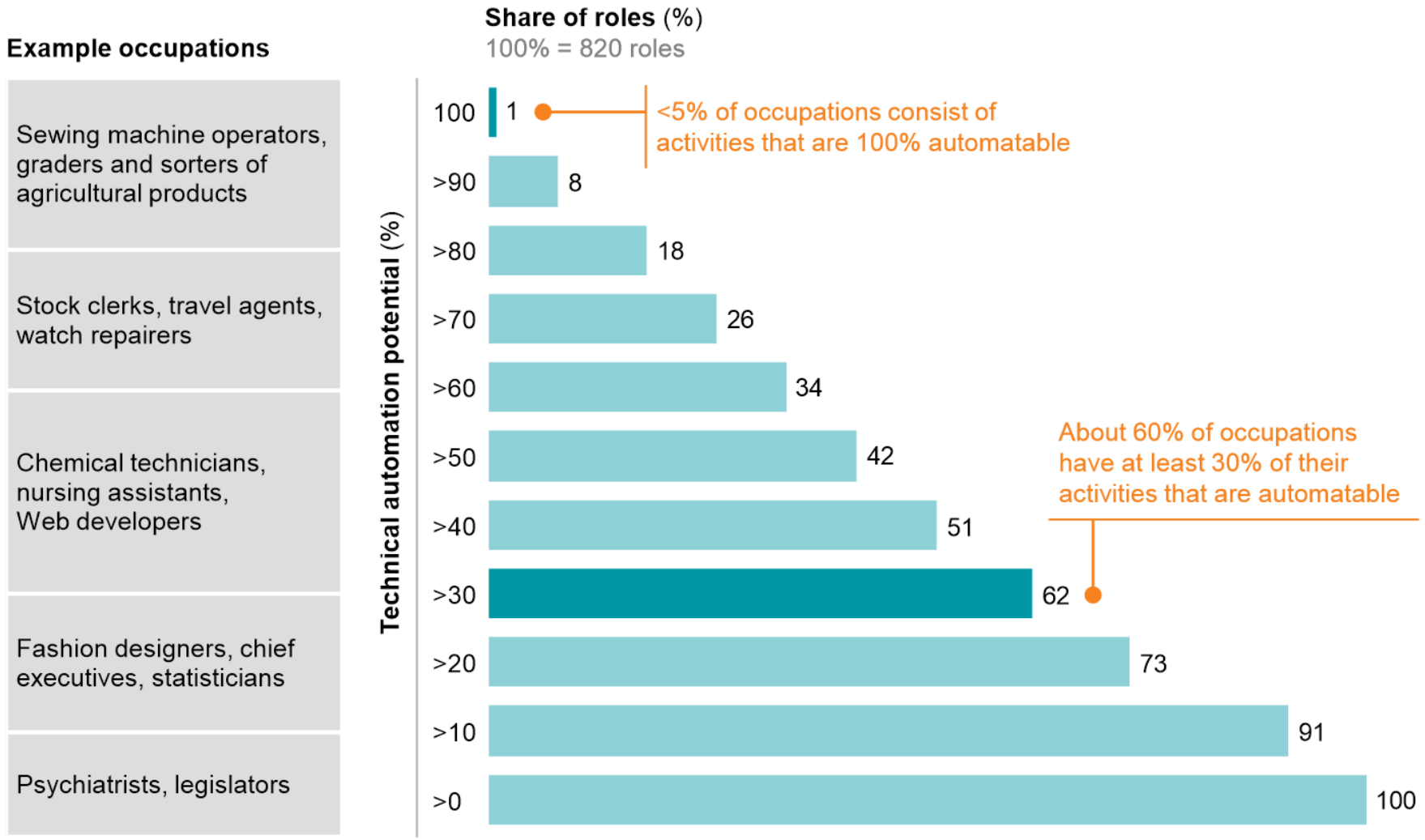
France Bans Judge Analytics, 5 Years In Prison For Rule Breakers

4th June 2019 artificiallawyer Litigation Prediction 52



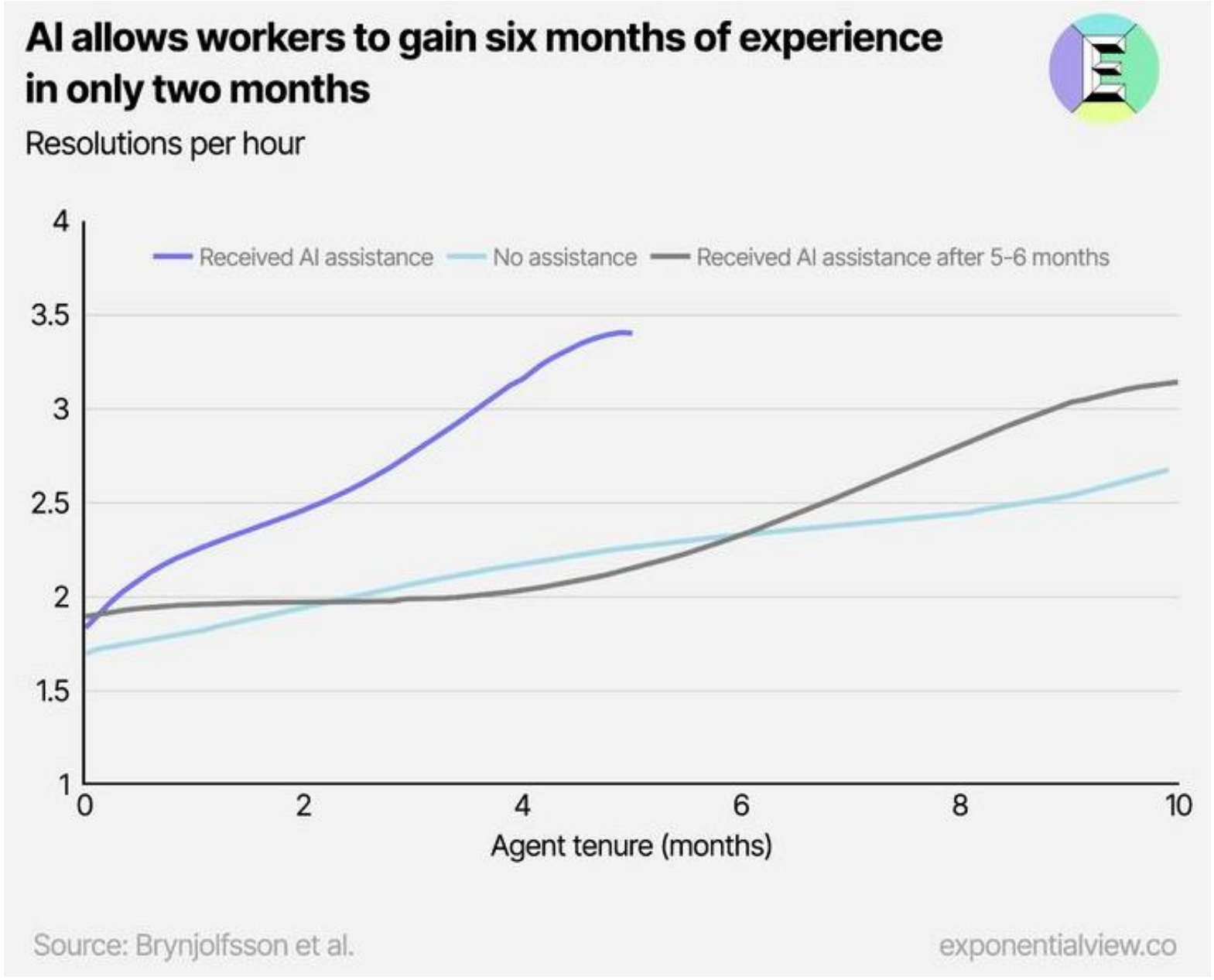
https://www.artificiallawyer.com/2019/06/04/france-bans-judge-analytics-5-years-in-prison-for-rule-breakers/

Automation potential based on demonstrated technology of occupation titles in the United States (cumulative)[1]



1 We define automation potential according to the work activities that can be automated by adapting currently demonstrated technology.

SOURCE: US Bureau of Labor Statistics; McKinsey Global Institute analysis

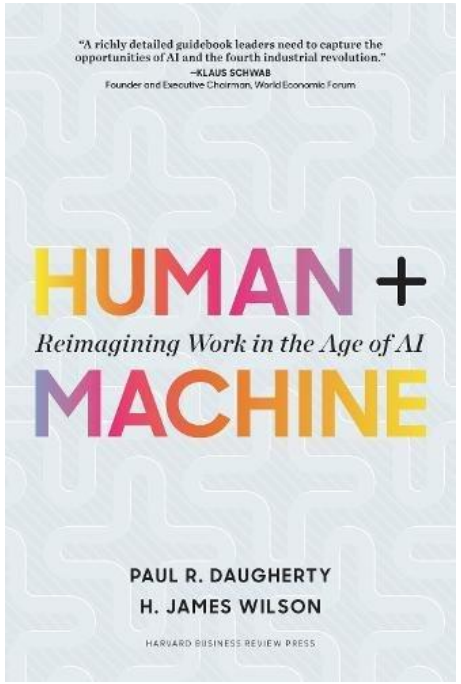


https://www.exponentialview.co/p/generative-intelligence-can-propel#details

Generative AI at Work, E. Brynjolfsson, D. Li & L. Raymond, 2023, DOI 10.3386/w31161

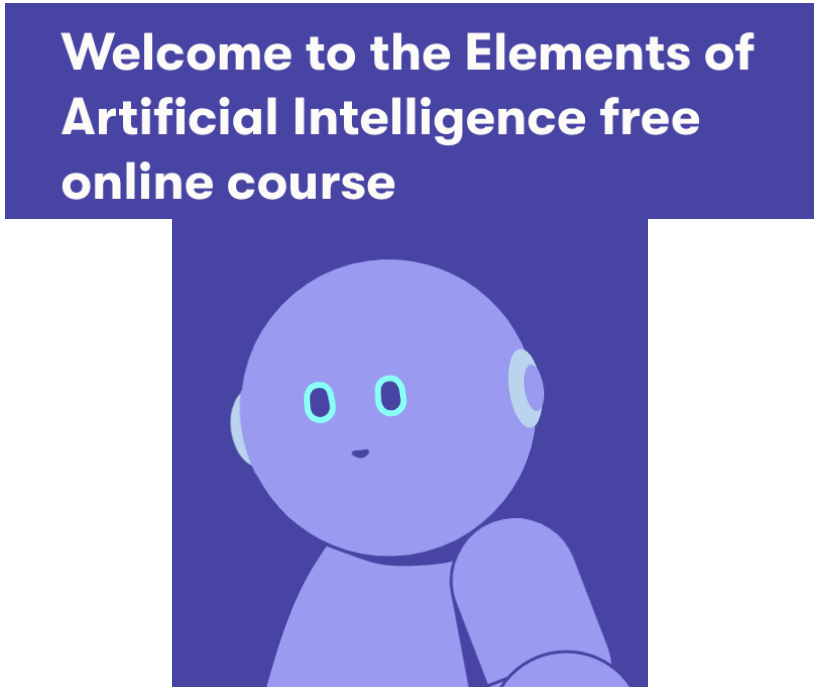
Other Components to Achieve Trustworthy AI

Humans + AI



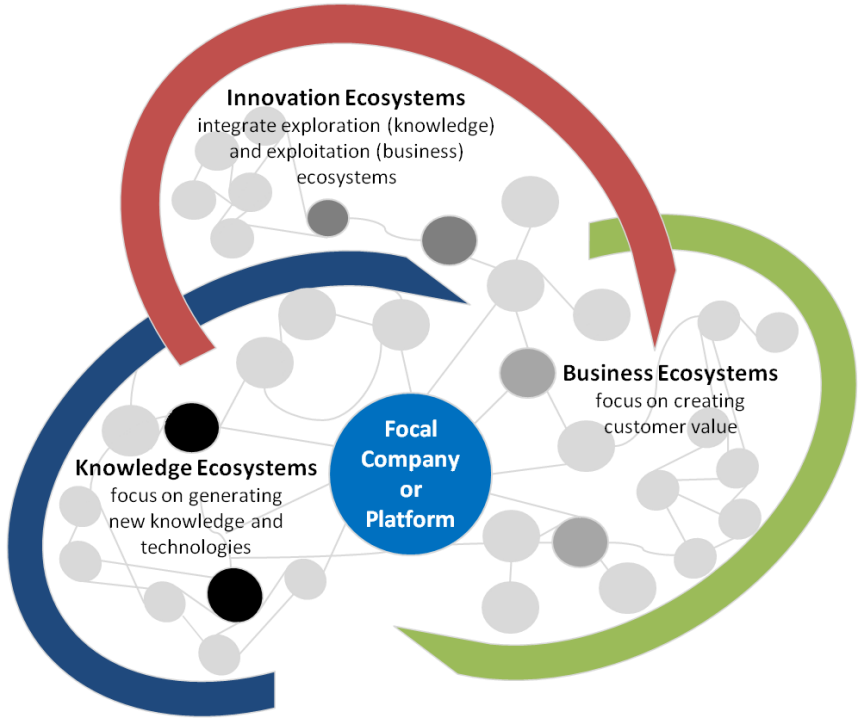
https://knowledge.wharton.upenn.edu/article/reimagining-work-age-ai/

Education



https://elementsofai.se

Ecosystems



https://timreview.ca/article/919

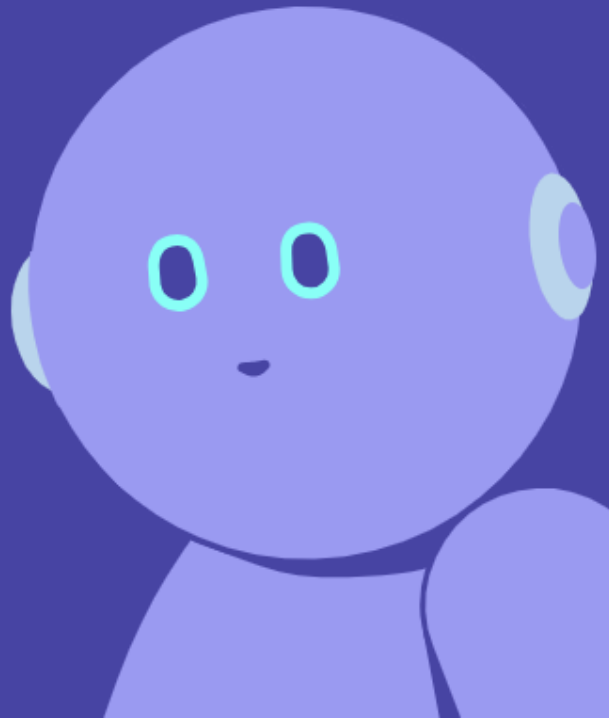
Elements of AI

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

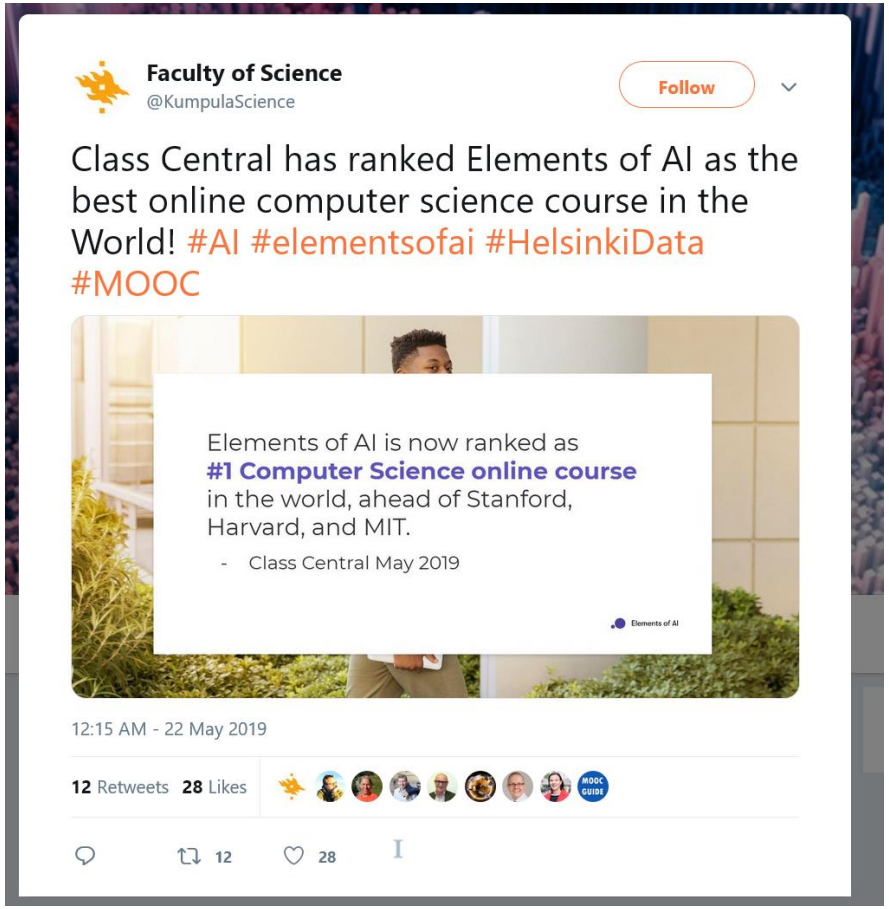
https://www.elementsofai.se/

Swedish launch funded by

VINNOVA
Sweden's Innovation Agency

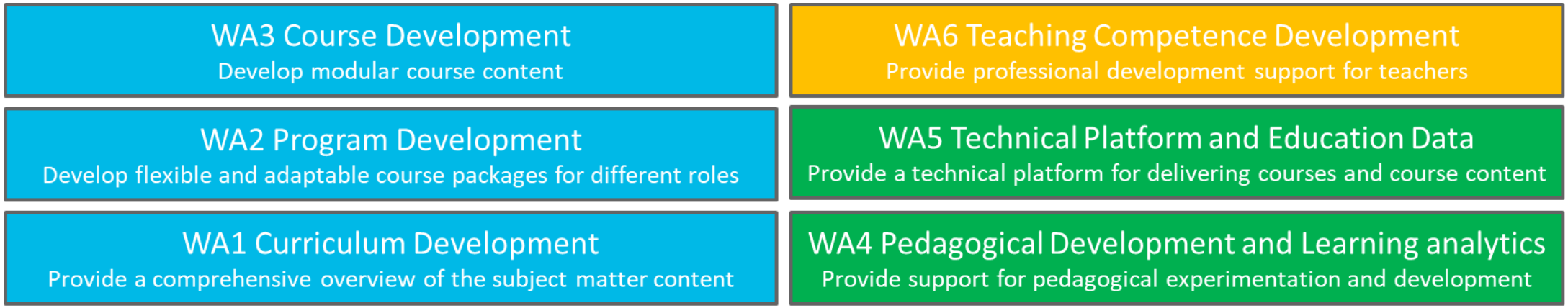
Elements of AI – Results so far

- Worldwide
  - > 1 115 000 signups
  - > 138 500 completed
- Sweden
  - > 47 500 signups
  - > 11 000 completed
  - > 11 500 signed up for the English course with Sweden as their country
  - > 7 000 have received university credits for the course

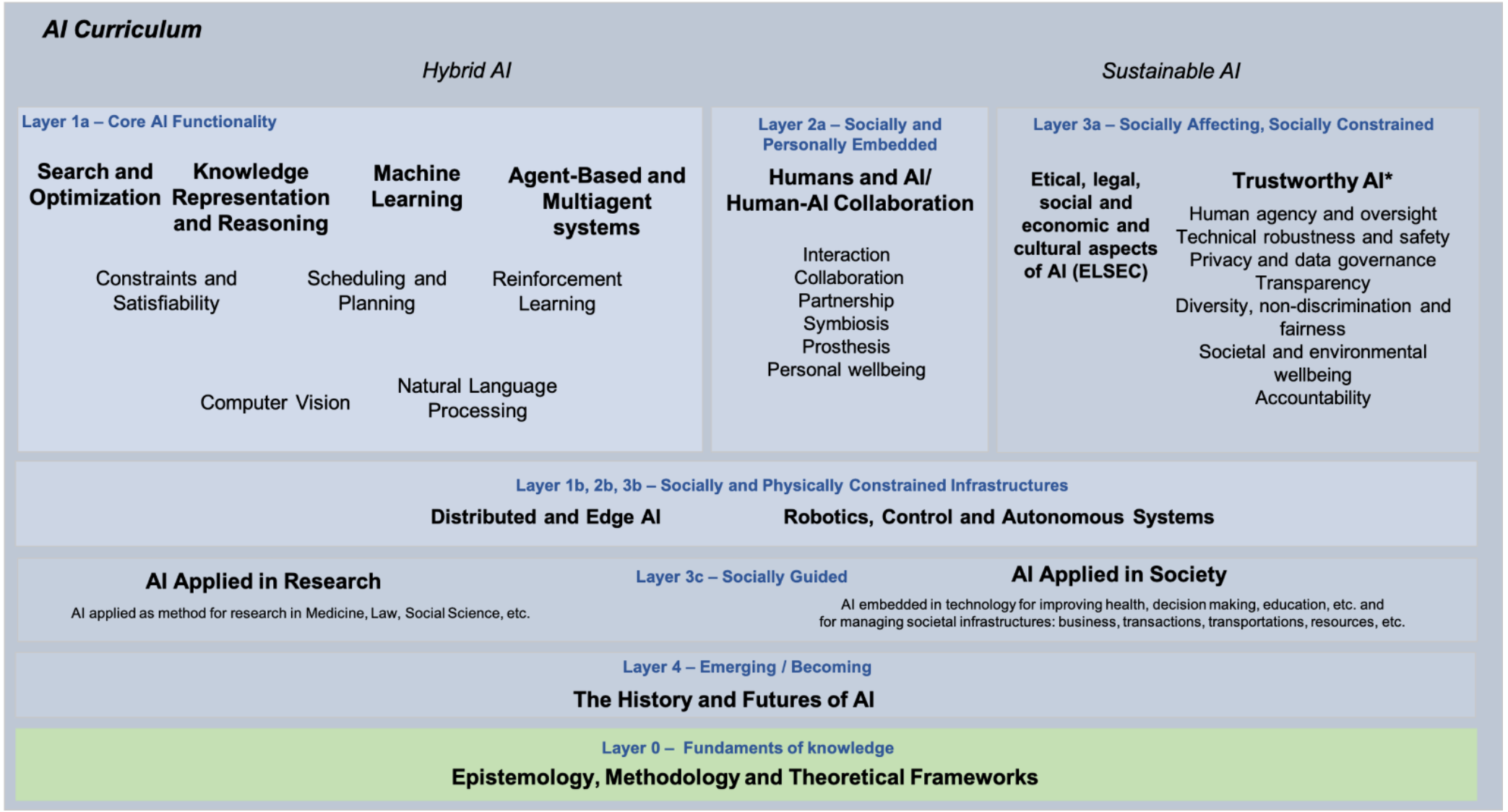


WALLENBERG AI AND TRANSFORMATIVE TECHNOLOGIES EDUCATION DEVELOPMENT PROGRAM

- Wallenberg AI and Transformative Technologies Education Development Program
- Purpose: significantly increase the capability and capacity of Swedish universities in providing timely, relevant, and scalable education in AI and other transformative technologies
- Objectives: 1) Provide educational foundations
  2) Scale-up the national educational capacity
  3) Scale-out education to disciplines and professions beyond the technical core
  4) Develop data-driven education and pedagogical transformation
- Work areas:



The WASP-ED AI Curriculum



LINKÖPINGS UNIVERSITET

H. Lindgren and F. Heintz. The WASP-ED AI Curriculum: A Holistic Curriculum for Artificial Intelligence. Proc. INETD.

WASP ED
WALLENBERG AI AND TRANSFORMATIVE TECHNOLOGIES EDUCATION DEVELOPMENT PROGRAM

AI and Computational Thinking – Two Sides of the same Coin

Fredrik Heintz, 2021. **The Computational Thinking and Artificial Intelligence Duality**. In *Computational Thinking Education in K-12: Artificial Intelligence Literacy and Physical Computing*. MIT Press.

AI Innovation, Competence and Research Ecosystem

TAILOR

AI SWEDEN

AI INNOVATION of Sweden

Elements of AI

AI Competence of Sweden

WASP ED

WASP—HS

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

CHALMERS UNIVERSITY OF TECHNOLOGY

KTH VETENSKAP OCH KONST

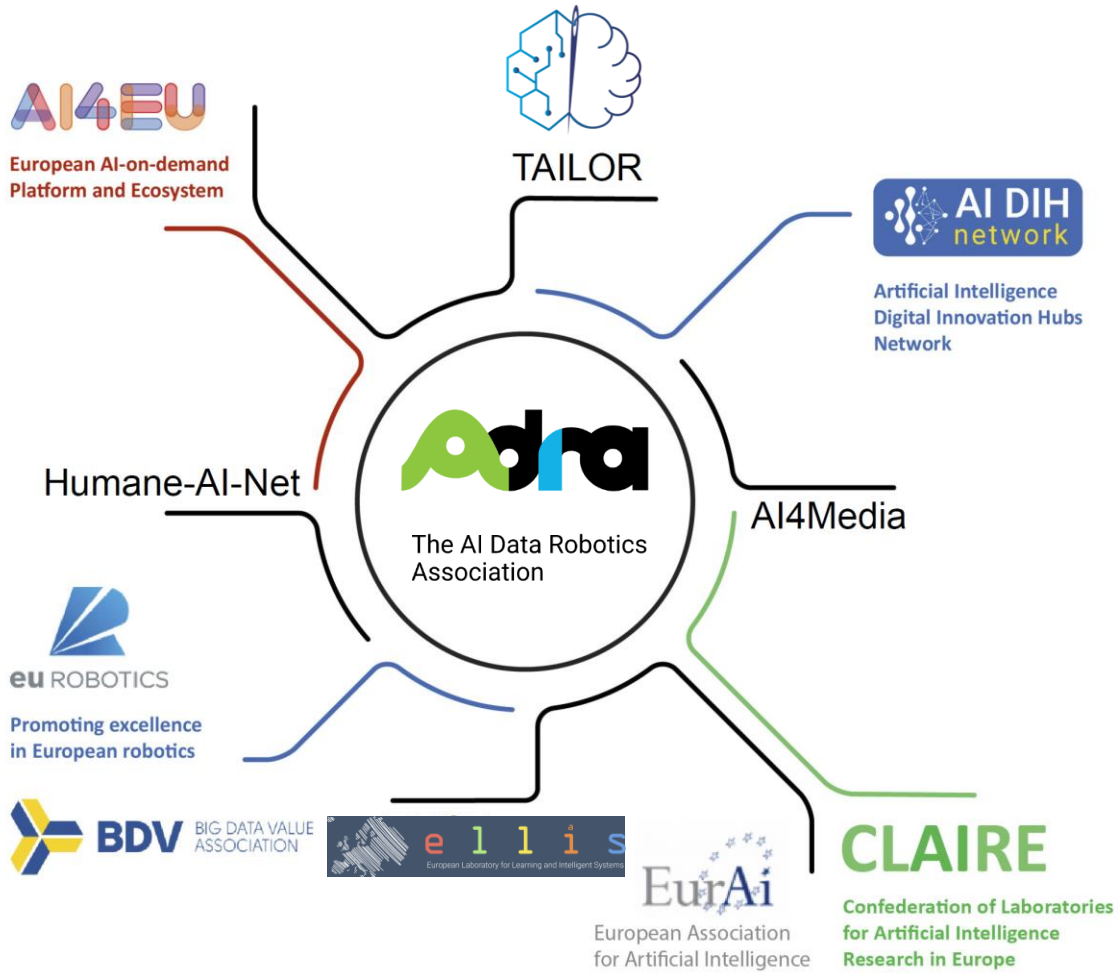
LINKÖPINGS UNIVERSITET

LUNDS UNIVERSITET

UMEÅ UNIVERSITET

UPPSALA UNIVERSITET

ÖREBRO UNIVERSITY



LINKÖPINGS UNIVERSITET

SWEDISH ARTIFICIAL INTELLIGENCE SOCIETY

PRIVATE INVESTMENT in AI by GEOGRAPHIC AREA, 2013–21

Source: NetBase Quid, 2021 | Chart: 2022 AI Index Report

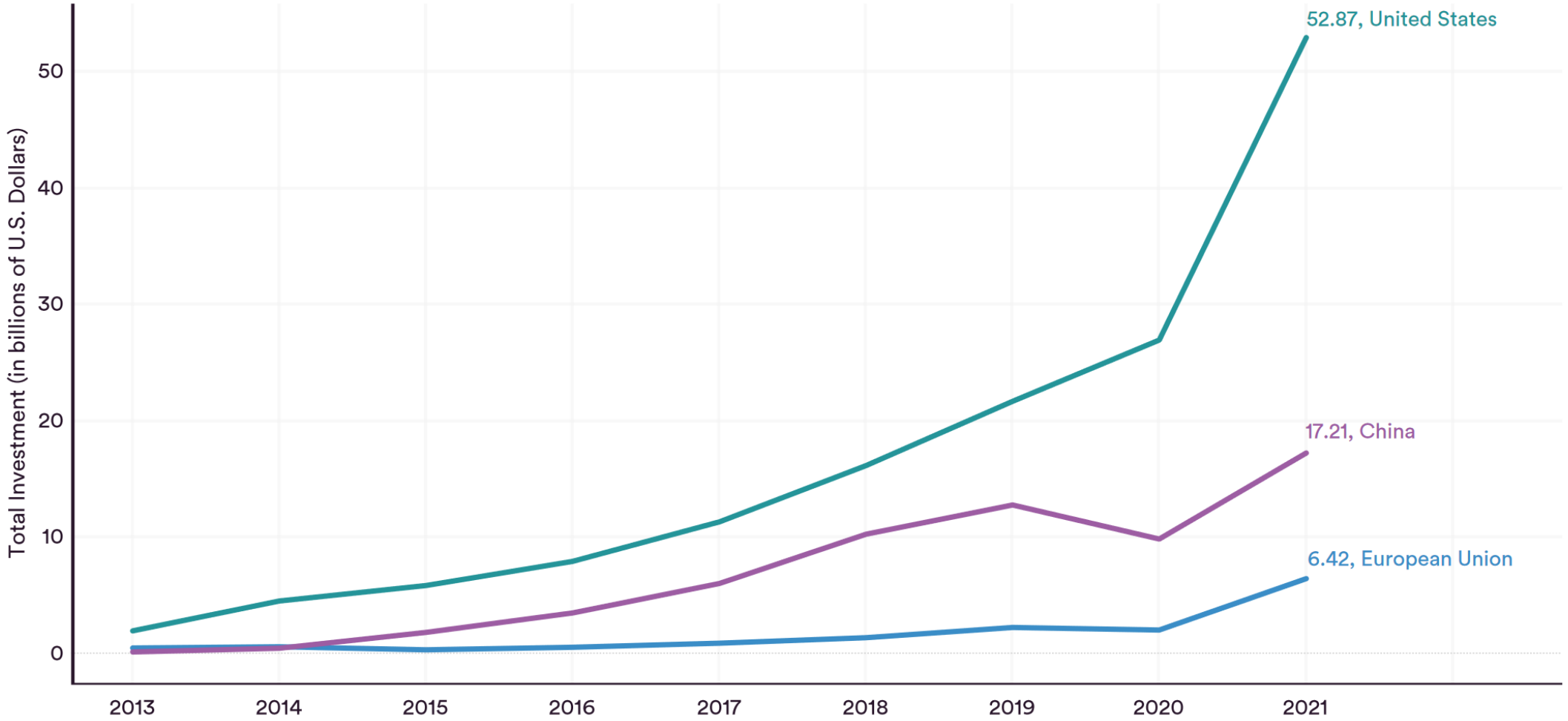
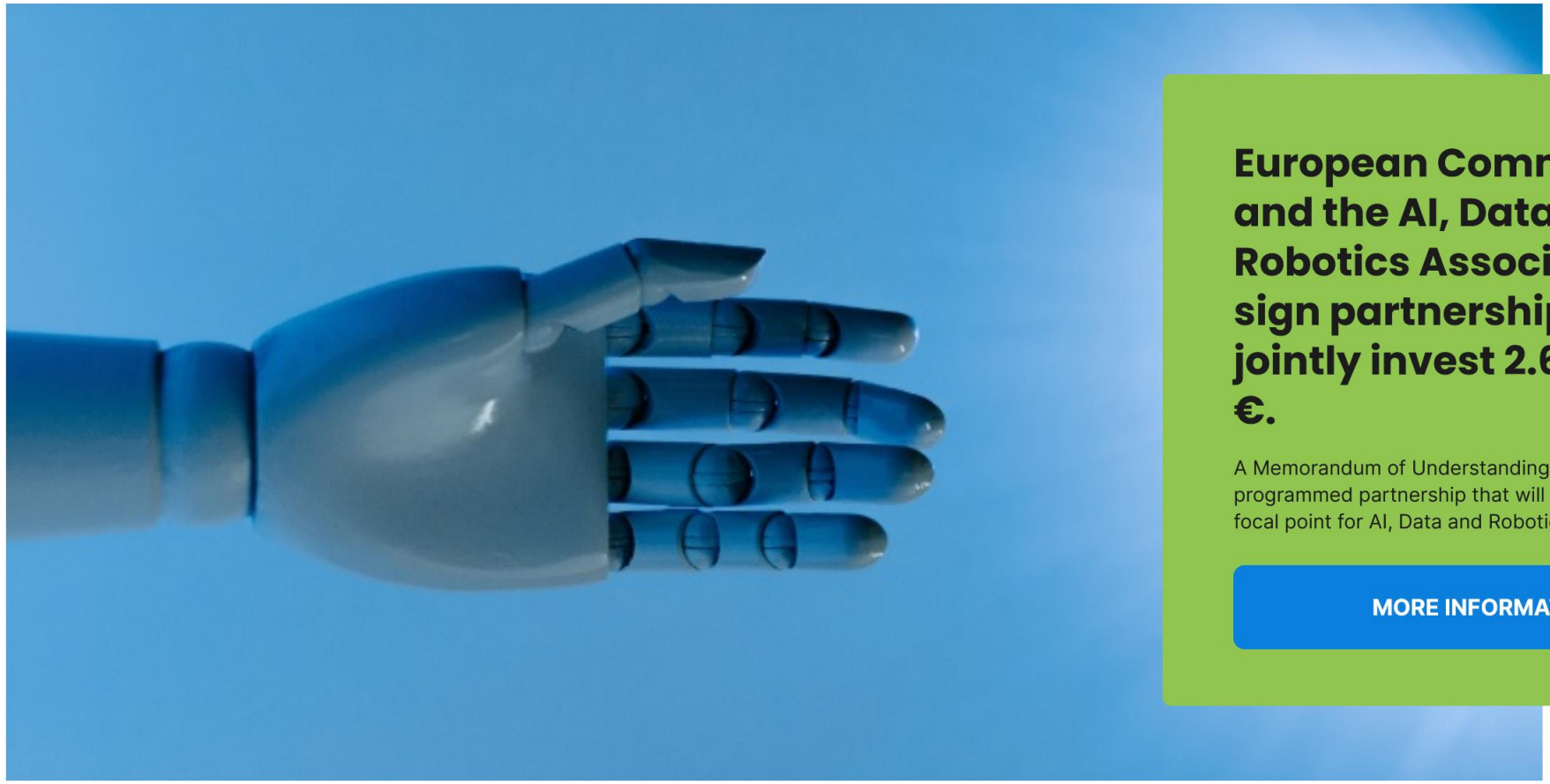


Figure 4.2.6

https://aiindex.stanford.edu/report/

The AI Data Robotics Association



European Commission and the AI, Data and Robotics Association sign partnership to jointly invest 2.6 Billion €.

A Memorandum of Understanding establishes the co-programmed partnership that will serve as European focal point for AI, Data and Robotics.

MORE INFORMATION

A join initiative by: BDV BIG DATA VALUE ASSOCIATION · CLAIRE · ellis · EurAi · eu ROBOTICS

https://adr-association.eu/

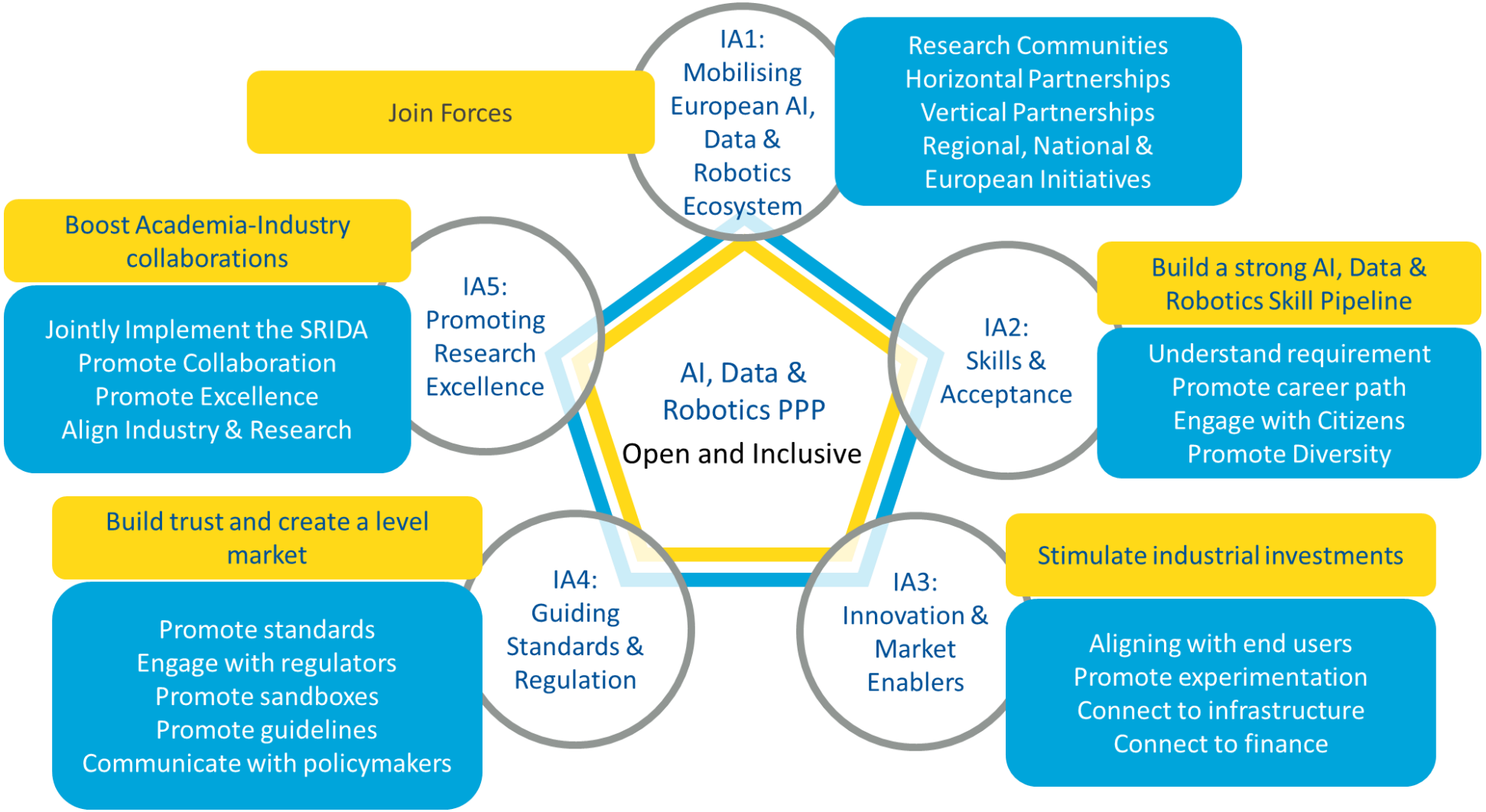
General Objectives of the ADR Partnership and Adra

Secure **European's sovereignty over AI, Data and Robotics** technologies and knowhow

Establish **European leadership in AI, Data and Robotics** technologies with high socio-economic and environmental impact

Reinforce a **strong and global competitive position of Europe** in AI, Data and Robotics

Implementing the Partnerships: Key IAs



Adra Position Paper – Strategic Directions High-Level

- **Trustworthy ADR technology made in Europe in compliance with the regulation** including the AI Act, the Data Act, and the Data Governance Act. Meeting regulation with innovation.
- **European strategic autonomy in ADR technology and the use of ADR technology to support strategic autonomy in other areas**, e.g., to optimize production cost and relocate production to EC.
- **Increasing the resilience of our society to crisis, both natural and man-made**. Improved preparedness as well as rapid, fast, and efficient response in catastrophic situations. Security and cybersecurity.
- **Green deal, sustainable society, zero carbon emission**. Operation, maintenance, and inspection of the circular economy and resource management.
- **Education on AI, Data and Robotics**, with a focus on scaling-up educational capacity and scaling-out education to other professions and subjects.

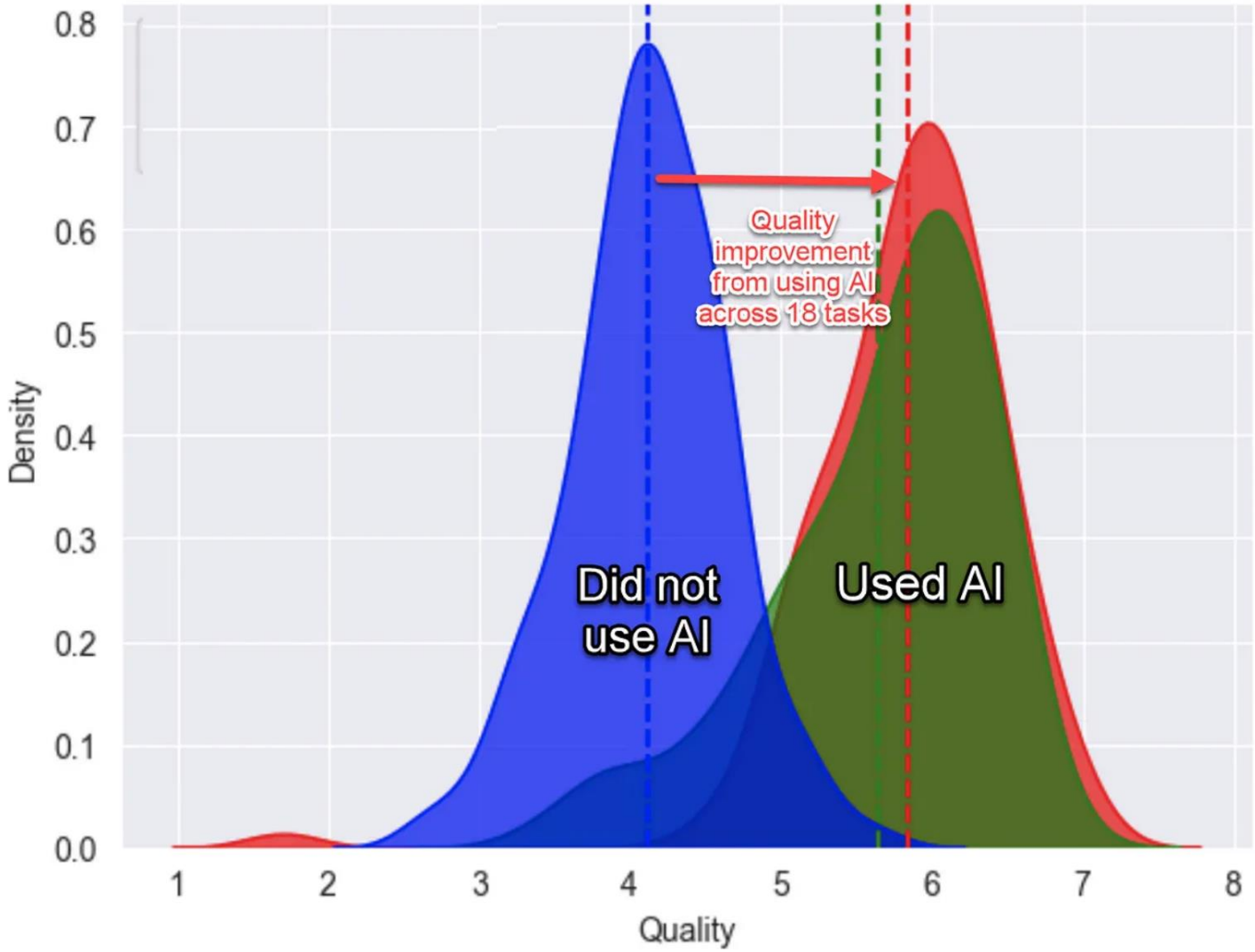
Adra Position Paper – Strategic Directions Technical

- **Large-scale general purpose/versatile generative ADR technology**. For example, open Large scale GDPR compliant European language models handling both language and cultural differences in Europe.
- **Large-scale complex ADR testbeds together with end-users** for example in healthcare, food production, transportation, energy, or smart cities.
- **Multi-stakeholder development, verification, validation, and integration of automated decision making** in socio-technical systems both for public and private sector.
- **Collaborative autonomous systems** interacting with both the environment and people. This includes autonomous drones in controlled airspace, last mile delivery, and self-driving vehicles.
- **Metrics for measuring progress in ADR**, with a special emphasis on Trustworthy ADR technology.

Adra Big Ticket Items

- **Ground-breaking technological foundations** in ADR (autonomy, high-performance and predictability)
- **Effective and Trustworthy General-Purpose AI**
- Interoperable and integrated framework for **data and model ecosystems** (operations, governance, privacy & security)
- Next generation **smart embodied robotic systems** (soft robotics, autonomy, manipulation, configurability, human robot interaction/collaboration)
- Developing **ADR technology** for the **sciences** (from data to knowledge to understanding)
- Research, innovation and tools for **compliance** (Trust, privacy, security beyond compliance)

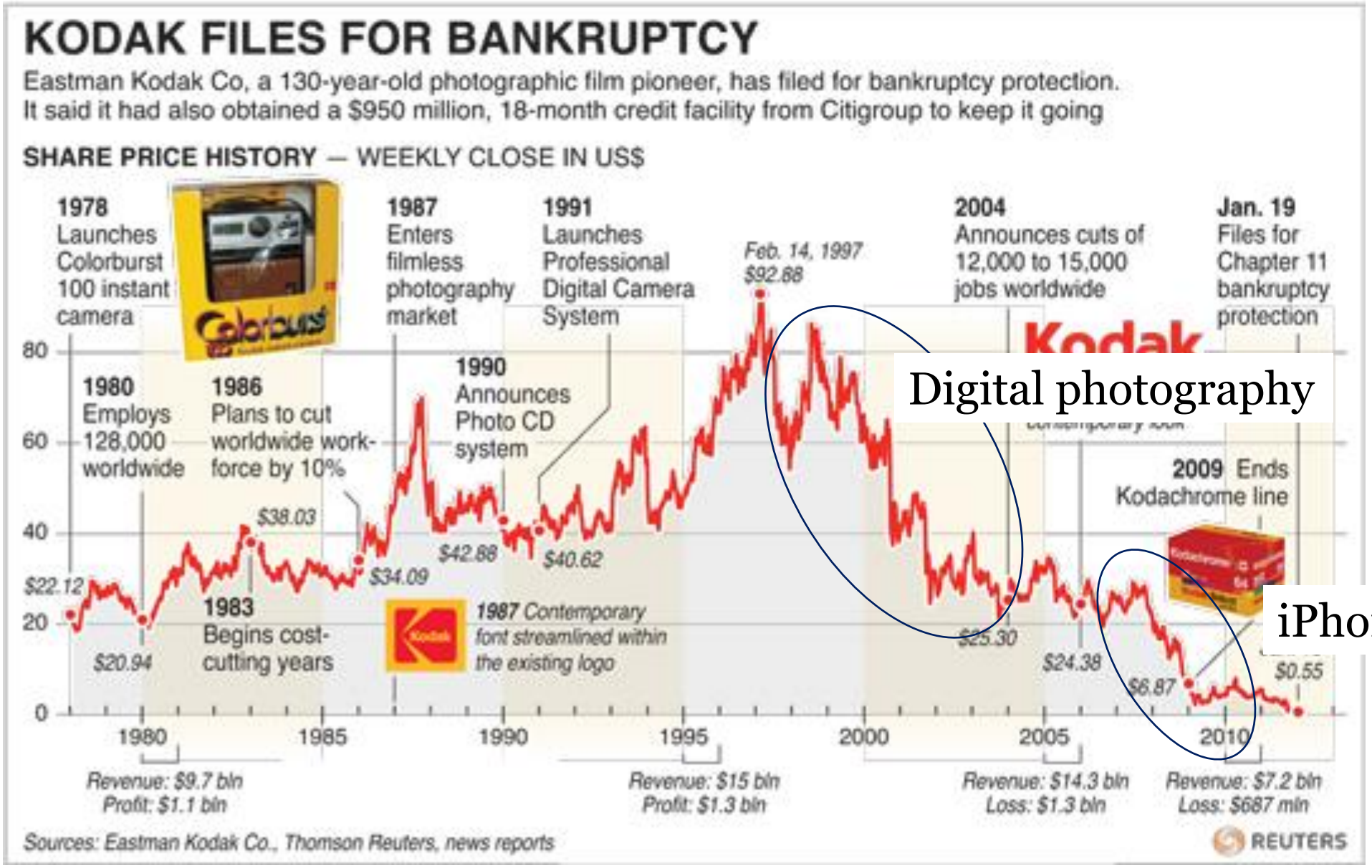
AI and Future of Work



Distribution of output quality across all the tasks. The blue group did not use AI, the green and red groups used AI, the red group got some additional training on how to use AI.

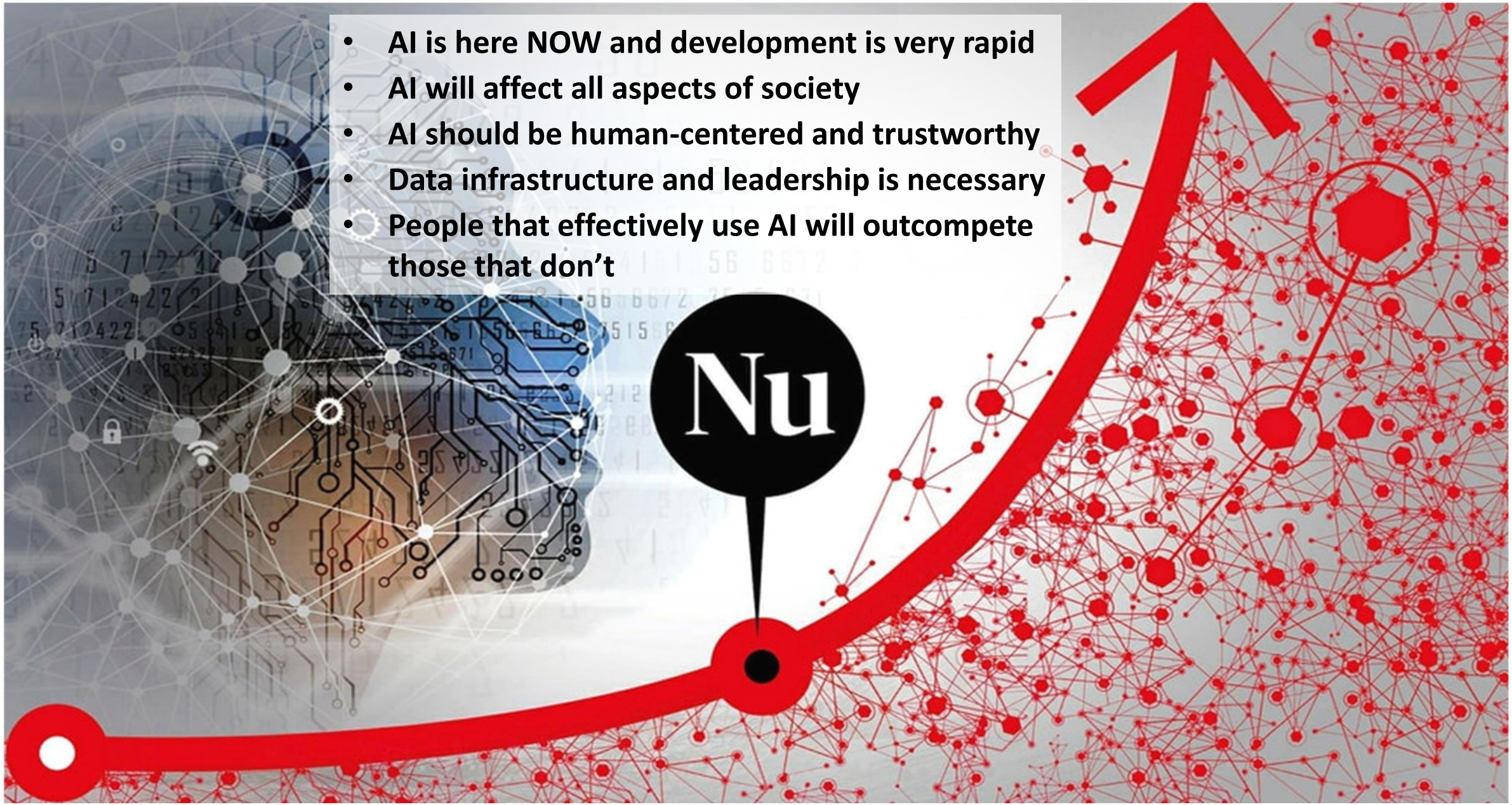
https://www.oneusefulthing.org/p/centaurs-and-cyborgs-on-the-jagged

HBS Working Paper 24-013 “Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality” by F. Dell'Acqua et al.



http://www.fpa-trends.com/article/how-manage-forecasting-risk

DAGENS NYHETER.



- AI is here NOW and development is very rapid
- AI will affect all aspects of society
- AI should be human-centered and trustworthy
- Data infrastructure and leadership is necessary
- People that effectively use AI will outcompete those that don’t

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