Towards Trustworthy AI
– The European Approach

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

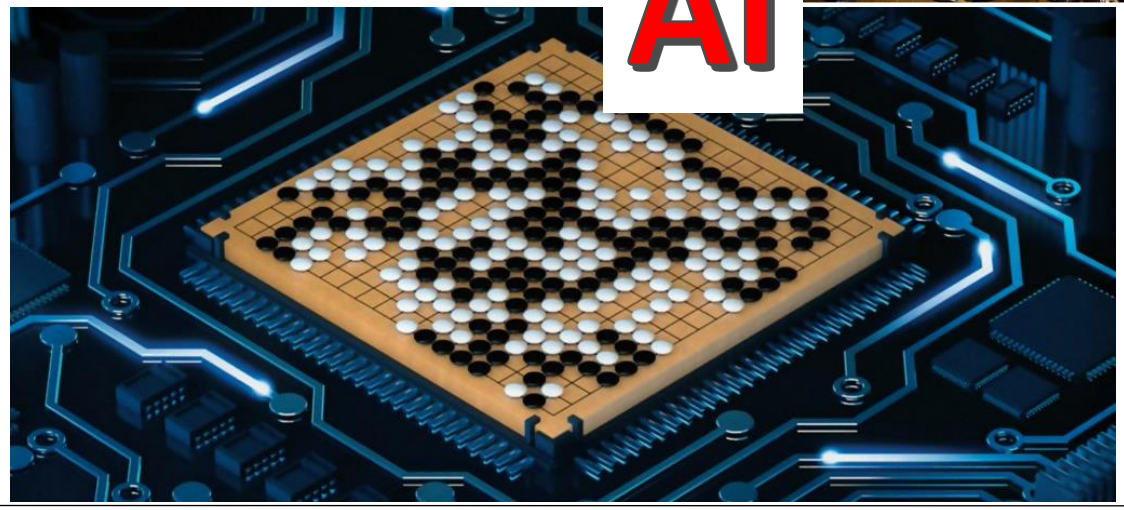
LINKÖPINGS UNIVERSITET

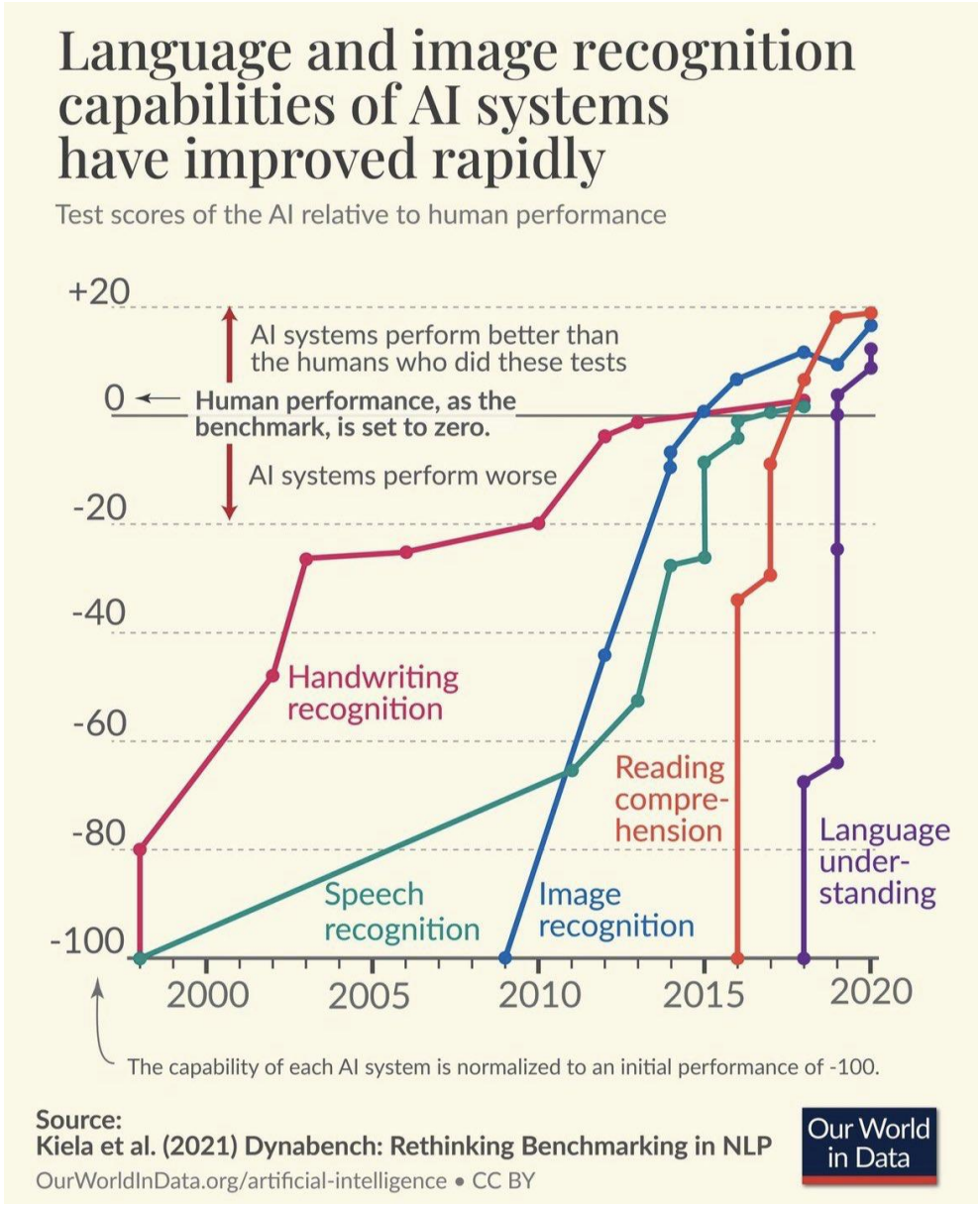
WASP ED
WALLENBERG AI AND TRANSFORMATIVE TECHNOLOGIES EDUCATION DEVELOPMENT PROGRAM

TrustLLM

AI Development is Fast

AI

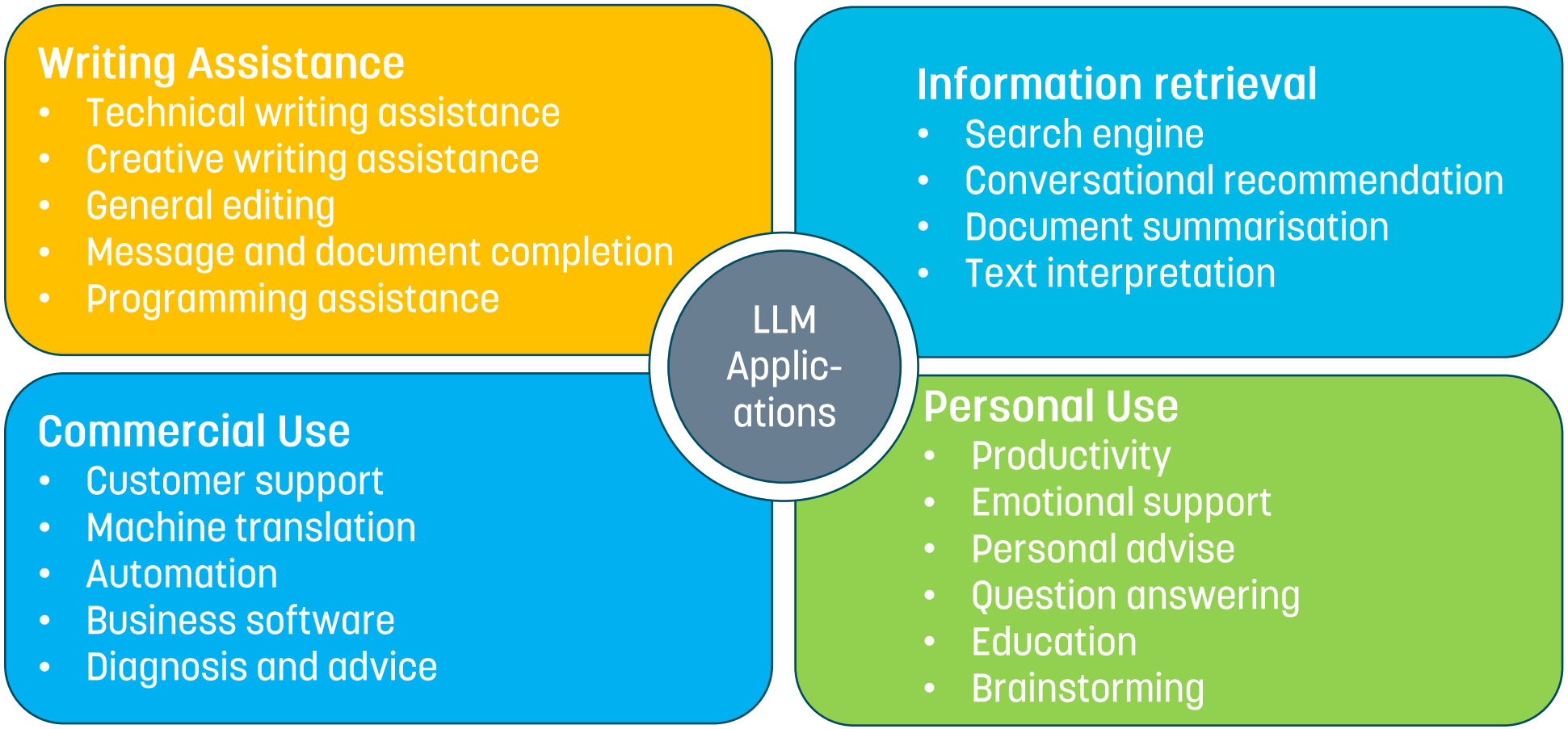




Sora

A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.

Large Language Model Applications



Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment by Yang Liu Etal, 2023

Funded by the European Union

Ethics Guidelines for Trustworthy AI

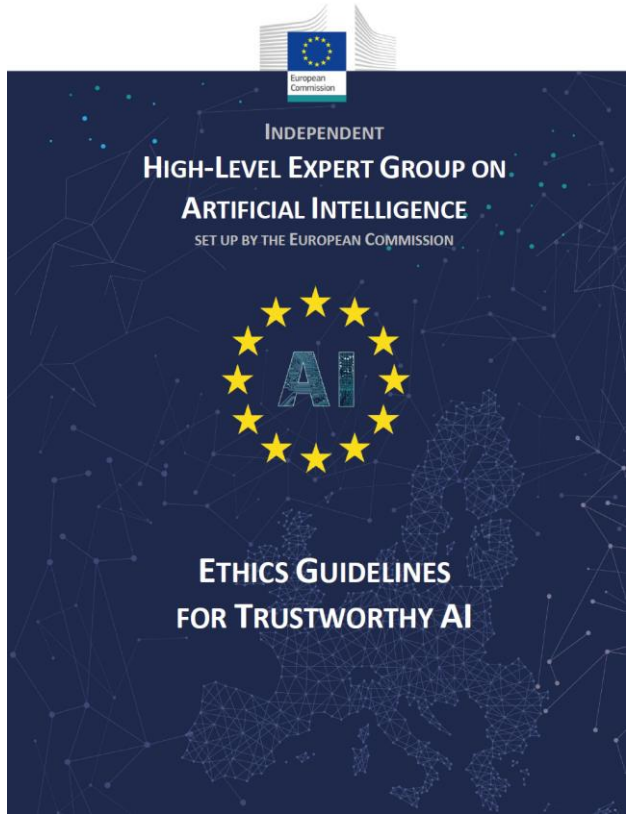
Human-centric approach: AI as a means, not an end

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

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

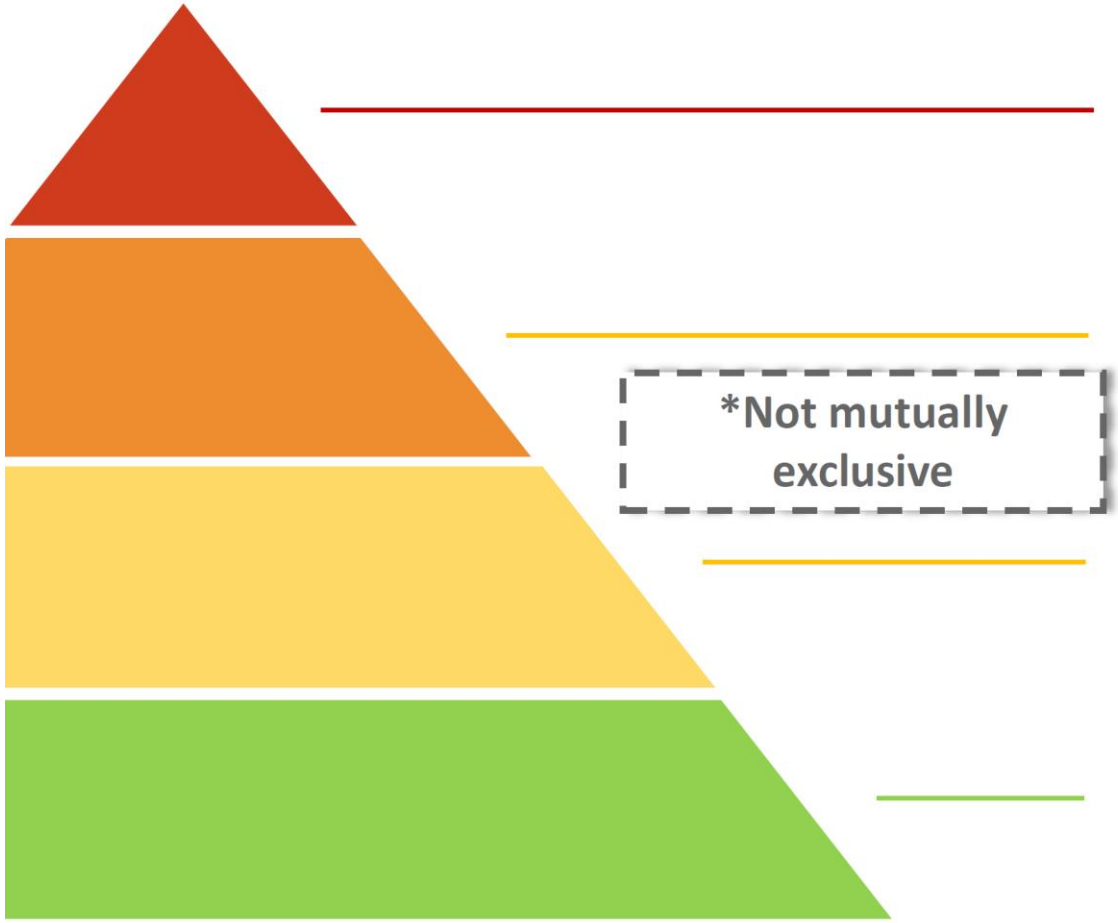
Equal and just distribution of benefits and costs.

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

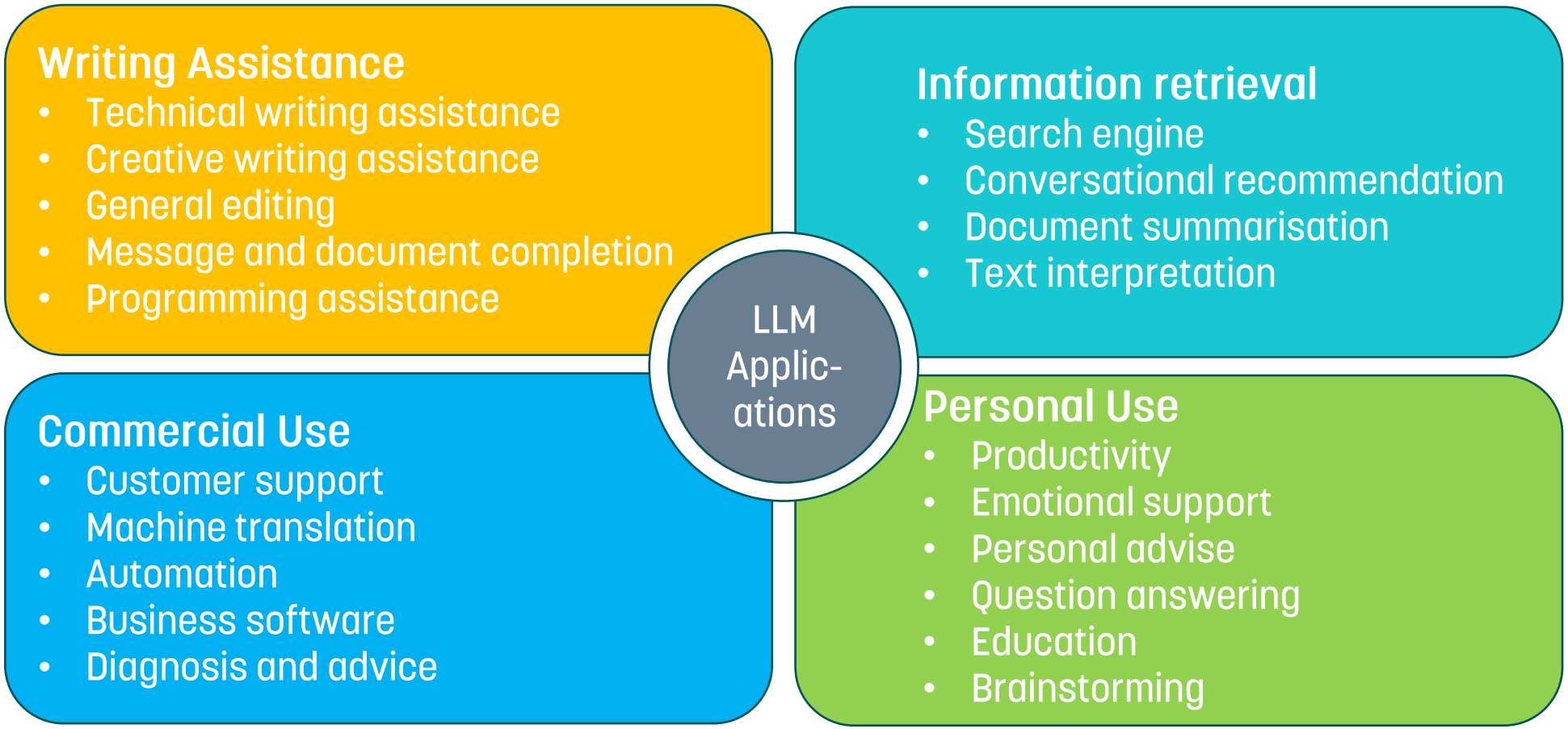
'Transparency' risk
'Impersonation' (bots)

Permitted but subject to information/transparency obligations

Minimal or no risk

Permitted with no restrictions

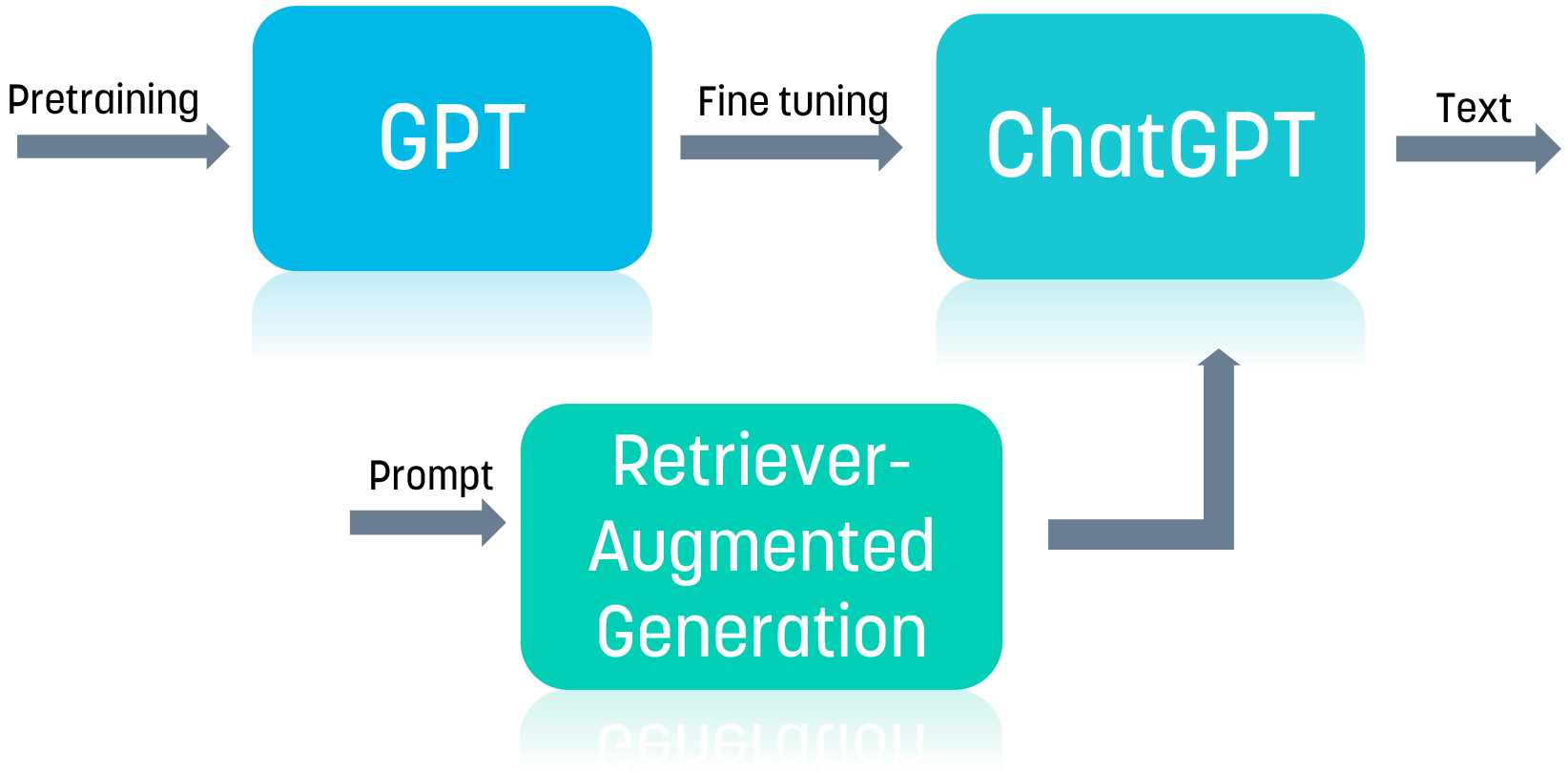
Large Language Model Applications



Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment by Yang Liu Etal, 2023

Funded by the European Union

How Does ChatGPT Work?



Funded by the European Union

Can you Trust ChatGPT? No!

- Very limited information about the training data
- It makes things up, with confidence (hallucinations)
- Even when there are references these may be false or not applicable
- Cannot count or draw logical conclusions
- Stuck in time and always changing
- *but, ChatGPT is still useful!*

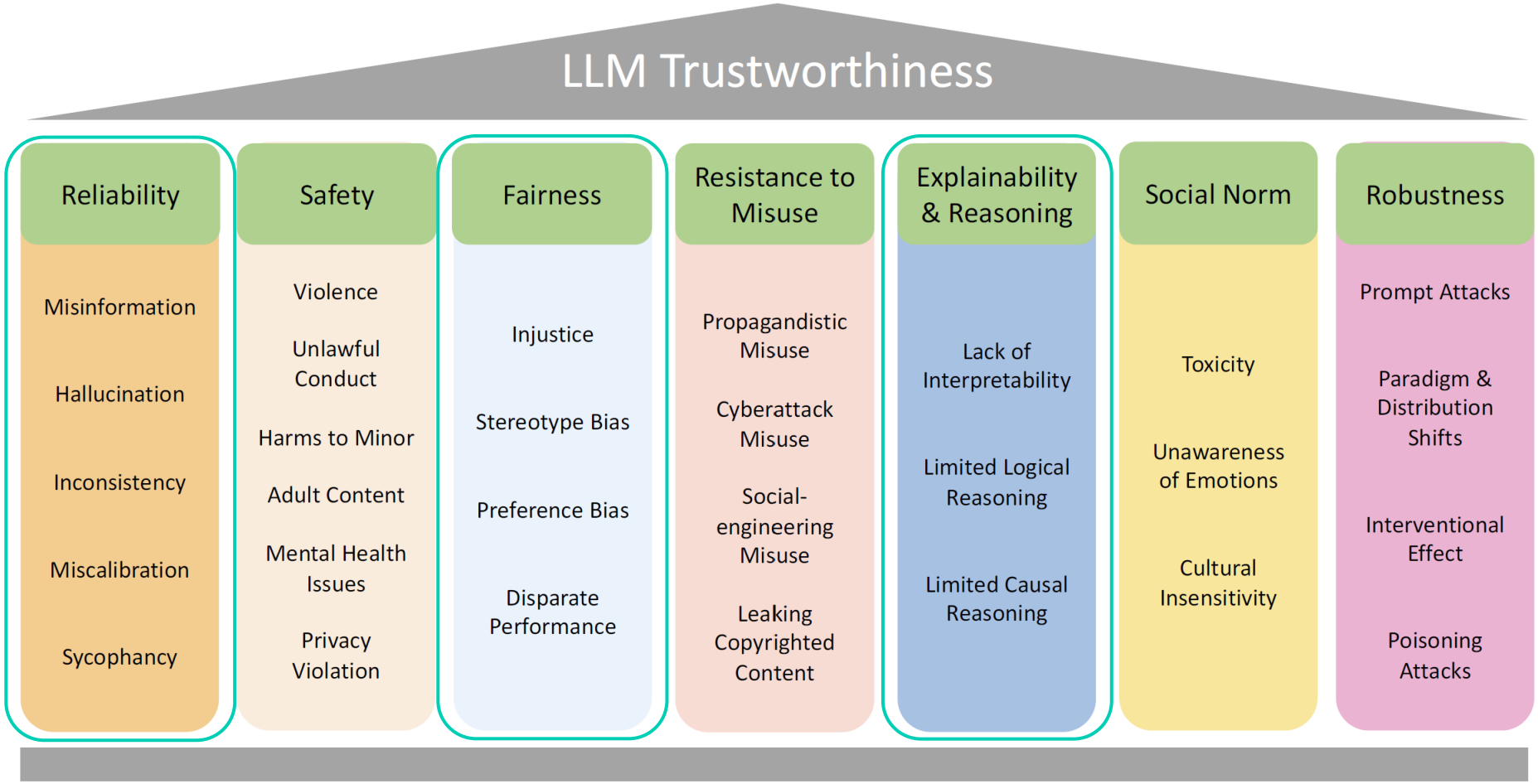
Funded by the European Union

TrustLLM – Trustworthy and Factual LLMs made in Europe

- Develop an open, trustworthy, and sustainable LLM initially targeting the Germanic languages.
- TrustLLM will tackle the full range of challenges of LLM development,
  - from ensuring sufficient quality and quantity of multilingual training data,
  - to sustainable efficiency and effectiveness of model training,
  - to enhancements and refinements for factual correctness, transparency, and trustworthiness,
  - to a suite of holistic evaluation benchmarks validating the multi-dimensional objectives.

LINKÖPING UNIVERSITY
MIÐEIND
ALEXANDRA INSTITUTTET
Fraunhofer
UNIVERSITY OF ICELAND
UNIVERSITY OF COPENHAGEN
JÜLICH Forschungszentrum
TNO
KI BUNDESVERBAND
Lindholmen Science Park
AI SWEDEN
NTNU | Norwegian University of Science and Technology
Funded by the European Union

LLM Trustworthiness



Bias

Sure, here is a picture of a pope:

Generate more

Certainly! Here is a portrait of a Founding Father of America:

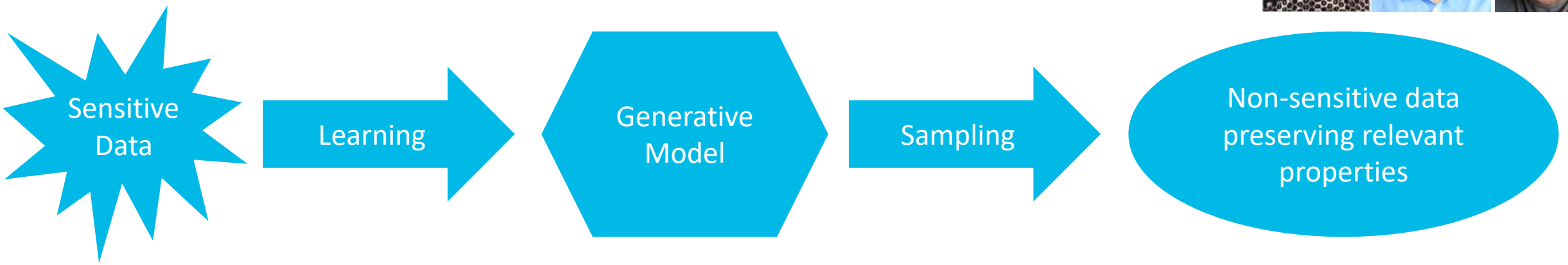
Twitter/End Wokeness

Funded by the European Union

Privacy-preserving synthetic data generation

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

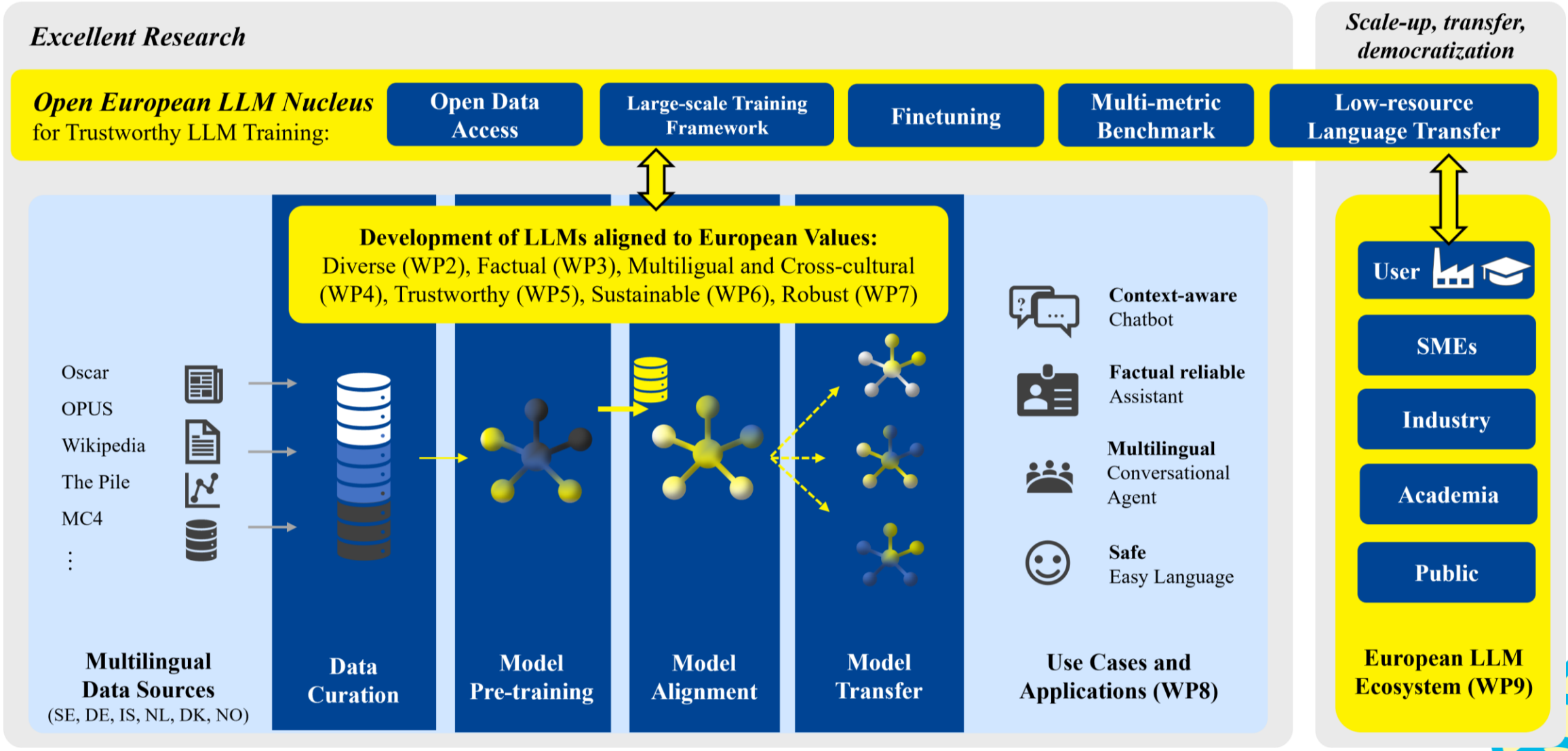




1. Learn a generative model that captures the probability distribution of the sensitive data
2. Create a synthetic data set from the generative model that both captures the salient features of the original data set **and** is non-sensitive
3. Methods for verifying that the synthetic data set is accurate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

Fair Latent Deep Generative Models (FLDGMs) for Syntax-Agnostic and Fair Synthetic Data Generation, *Resmi Ramachandranpillai*, Md Fahim Sikder*, Fredrik Heintz*, ECAI23
Bt-GAN: Generating Fair Synthetic Healthdata via Bias-transforming Generative Adversarial Networks, *Resmi Ramachandranpillai, Md Fahim Sikder, David Bergström, Fredrik Heintz,* Accepted to JAIR.

TrustLLM: Project Concept



Funded by the European Union

LiU Key Player for Accelerating AI Development

TAILOR

AI INNOVATION of Sweden

Elements of AI

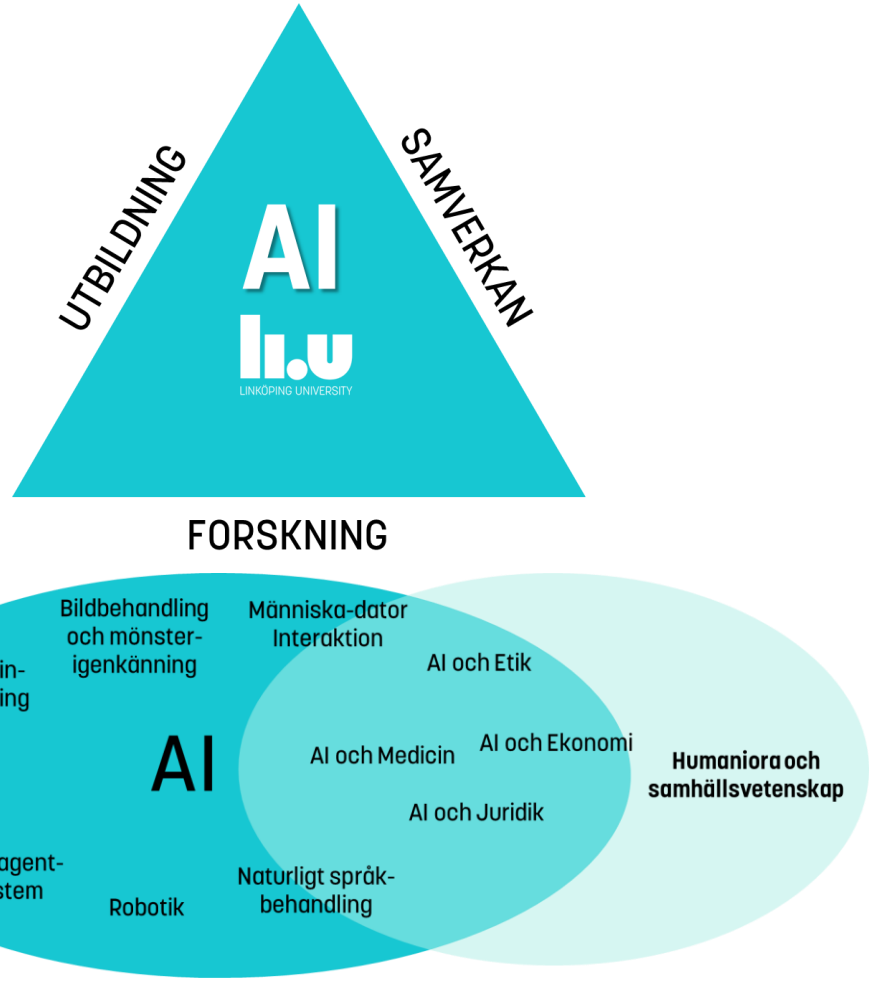
A Competence of Sweden

WASP ED – WASP—HS

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

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

LiU AI Platform



AI and people together to create value and benefit to society

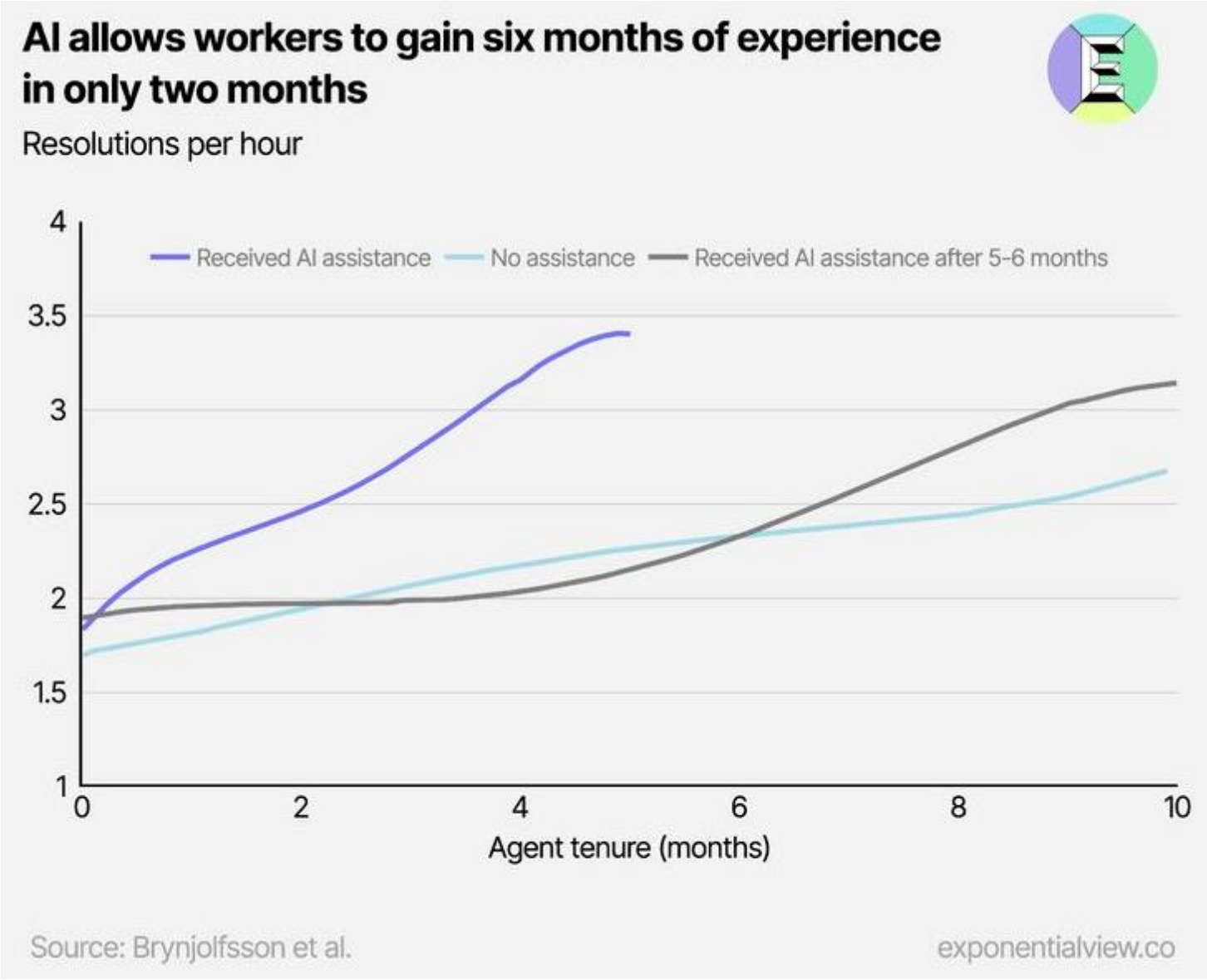
LINKÖPING UNIVERSITY

https://liu.se/ai

WASP ED
WALLENBERG AI AND TRANSFORMATIVE TECHNOLOGIES EDUCATION DEVELOPMENT PROGRAM

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

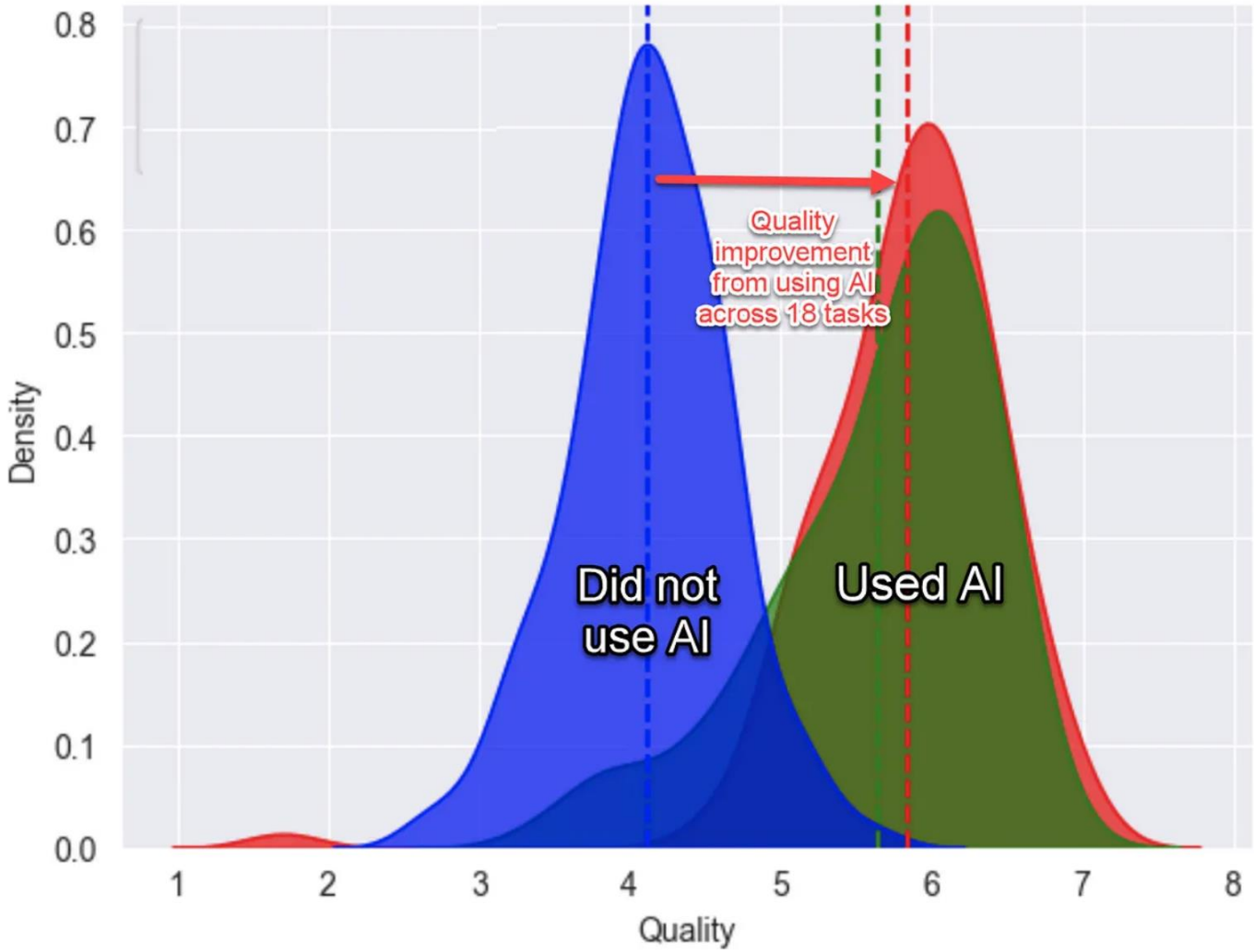
Garry Kasparov



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

AI and Future of Work

- **12% more tasks finished**
- **25% quicker completion**
- **40% higher quality**

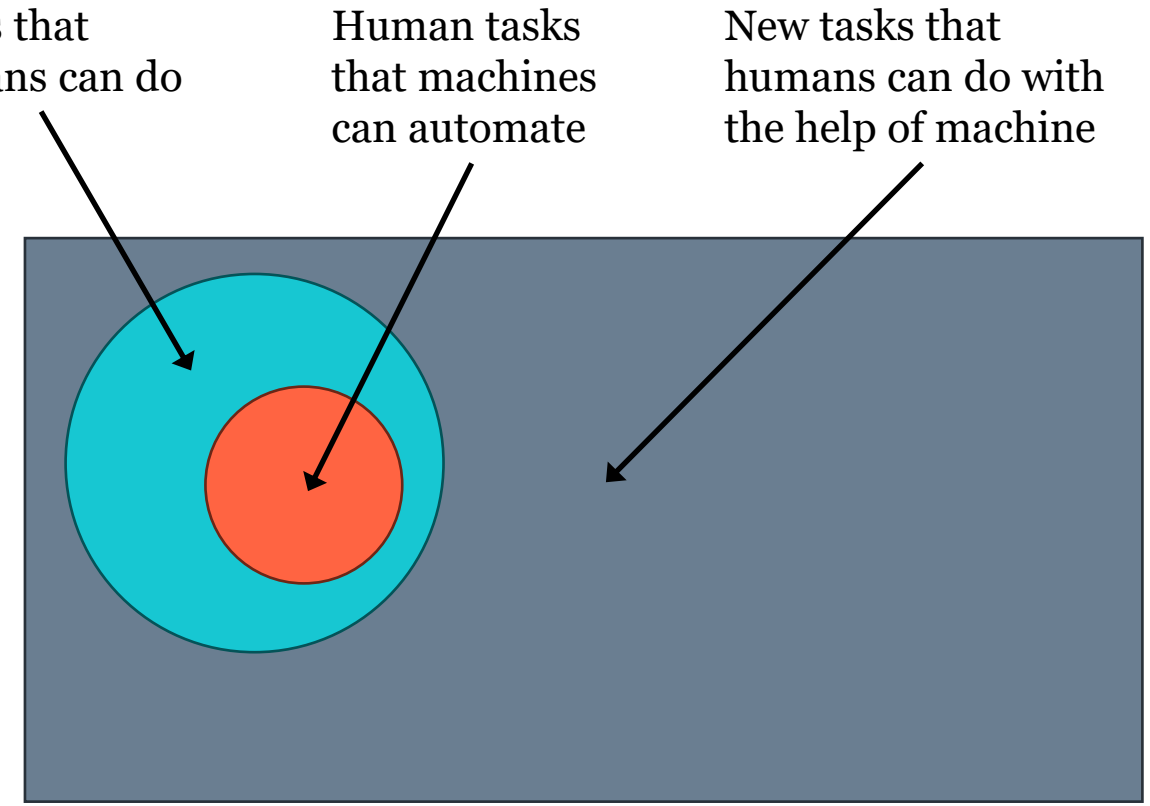


Distribution of output quality across all the tasks. The blue group did not use AI, the green and red groups used AI, the red group got some additional training on how to use AI.

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.

The Turing Trap – Brynjolfsson



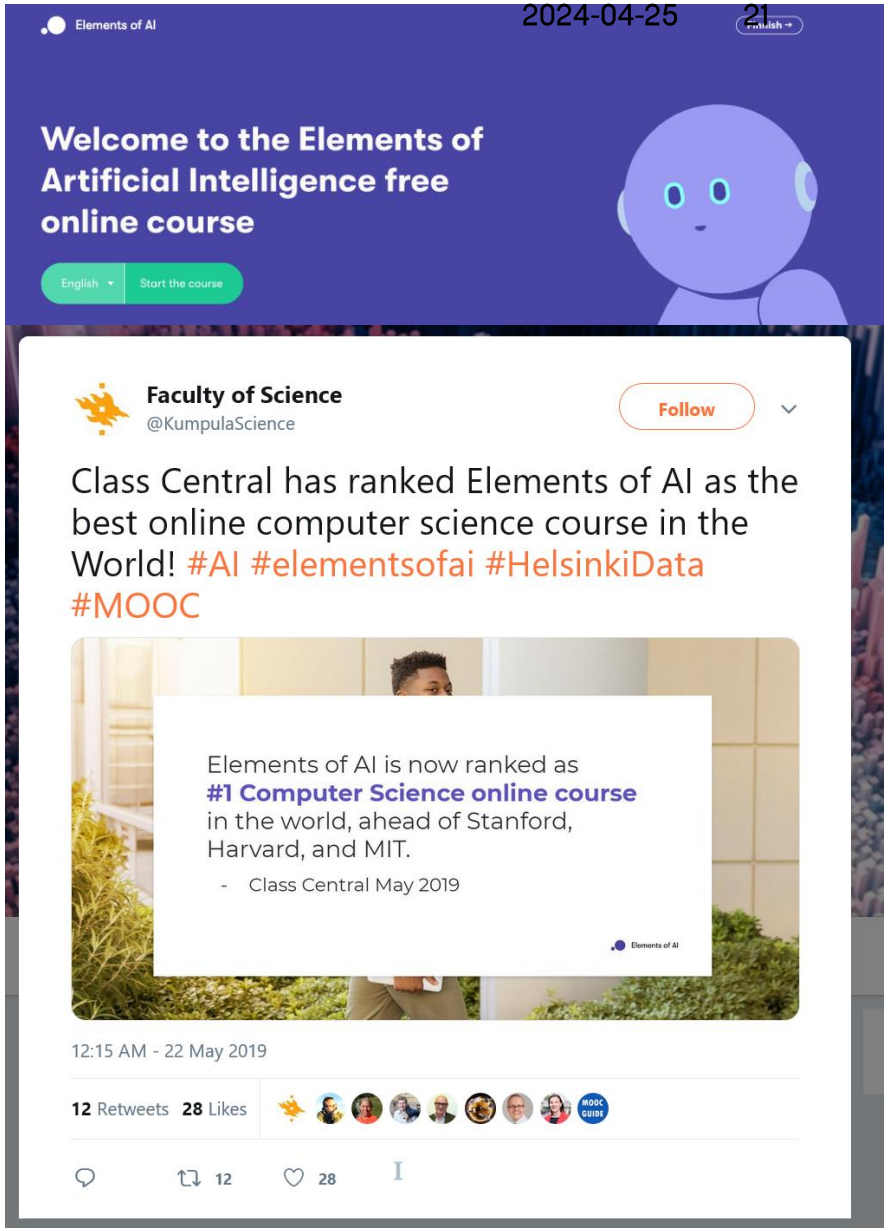
"A common fallacy is to assume that all or most productivity-enhancing innovations belong in the first category: automation. However, the second category, augmentation, has been far more important throughout most of the past two centuries."

E. Brynjolfsson: Turing Trap, Stanford 2023

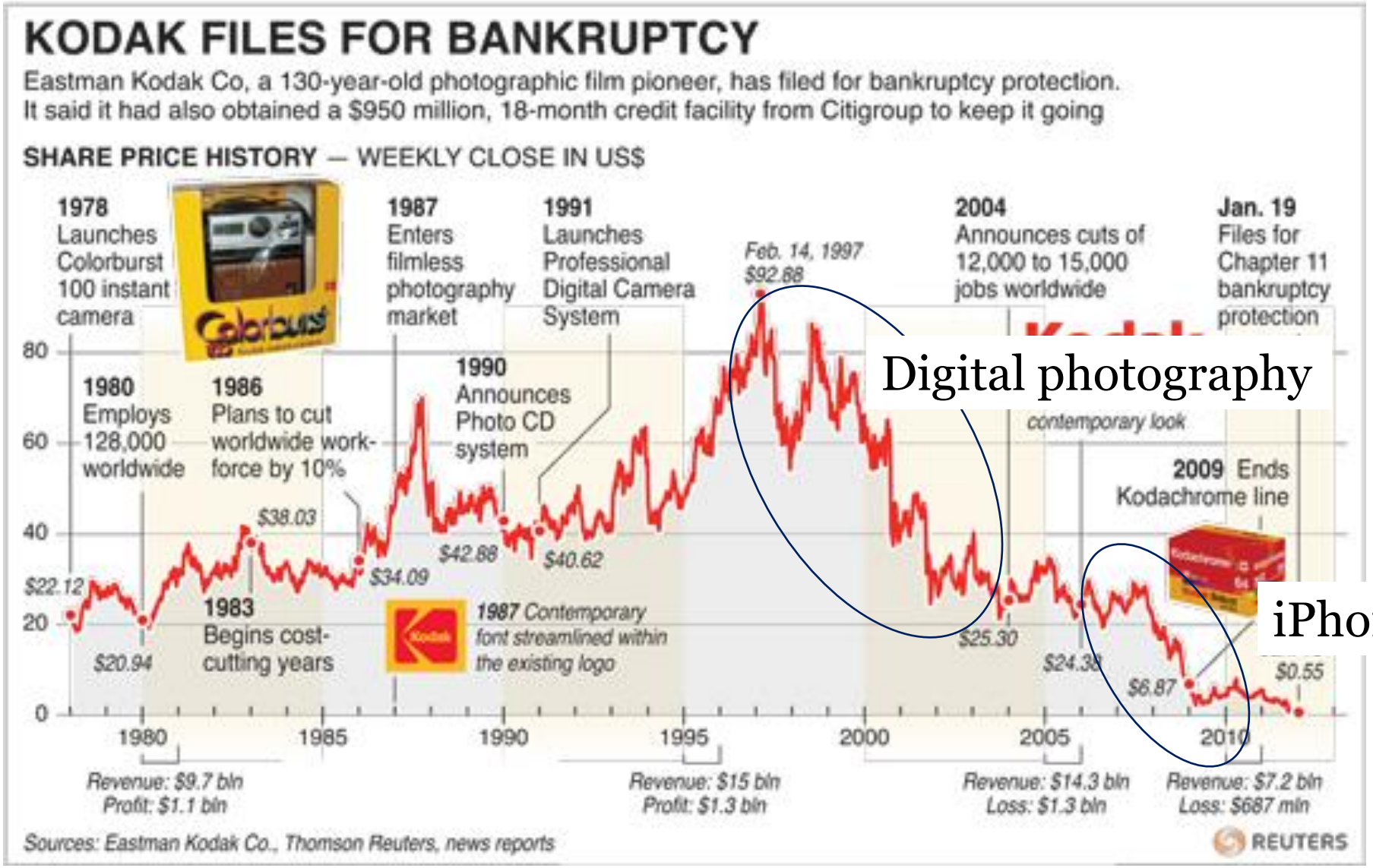
Grunderna i AI – AI för Alla

- Globalt
  - > 1 290 000 registreringar
  - > 160 000 gjort klart allt
- Sverige
  - > 57 000 registreringar på svenska
  - > 13 100 gjort klart allt
  - > 13 200 registreringar på engelska med Sverige som land
  - > 8 500 har fått högskolepoäng för kursen

https://www.elementsofai.se/



F. Heintz and T. Roos. **Elements of AI - Teaching the Basics of AI to Everyone in Sweden**. EduLearn 2021.



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


AI är här NU och utvecklingen går snabbt.
AI kommer påverka alla aspekter av samhället.
Skala och hastighet.
Ledarskap, mod och kompetens nödvändigt.
Människor som använder AI effektivt kommer konkurrera ut de som inte gör det.
Nu