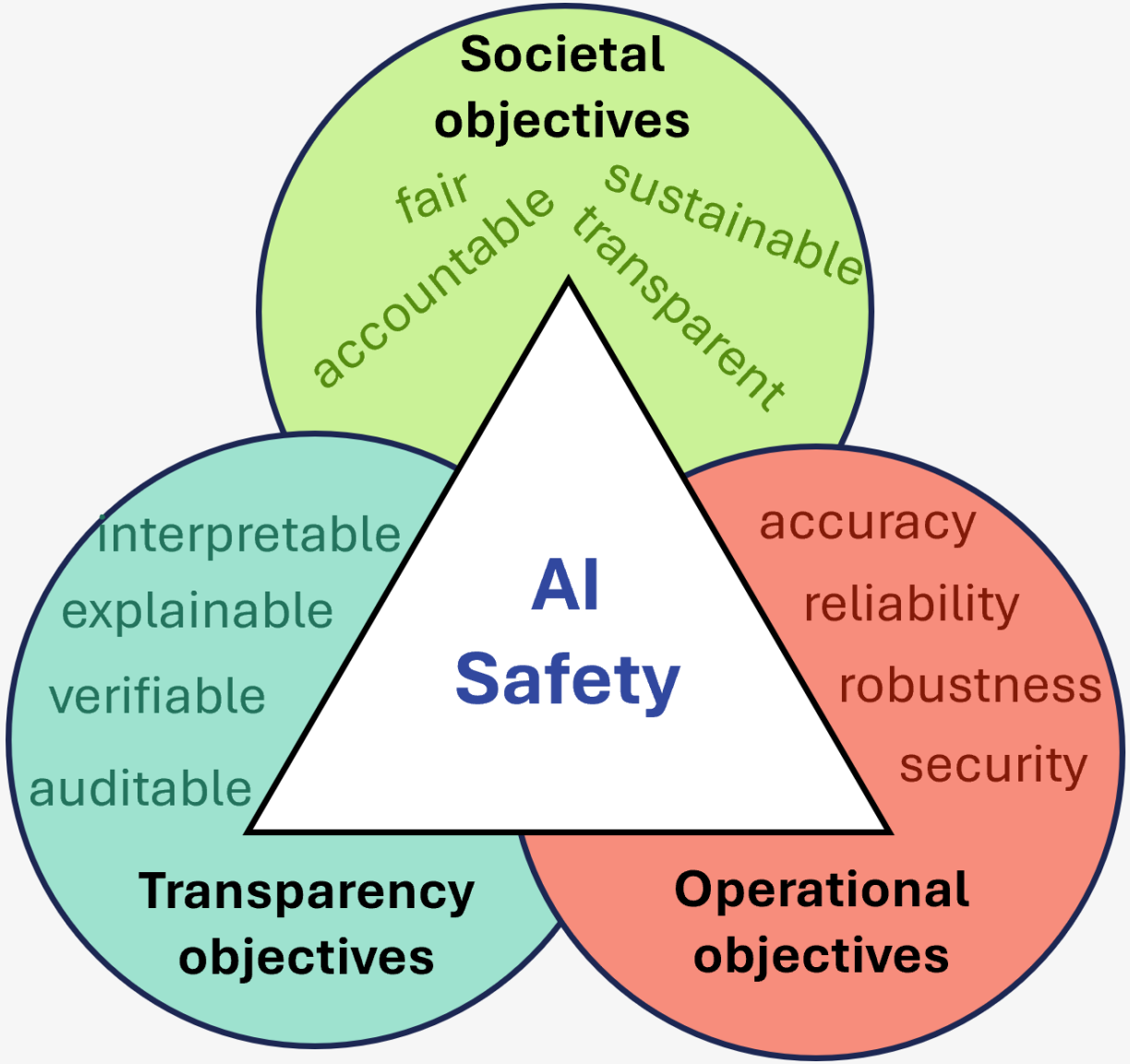


Artificial Intelligence Ethics

Dr Varun Ojha

Senior Lecturer in Artificial Intelligence

School of Computing

Newcastle University

varun.ojha@ncl.ac.uk

Artificial Intelligence

What is Artificial Intelligence?

A field of study that seeks to explain and emulate intelligent behaviour in terms of computational processes.

Schalkoff, 1990

The study of how to make computers do things at which, at the moment, people are better.

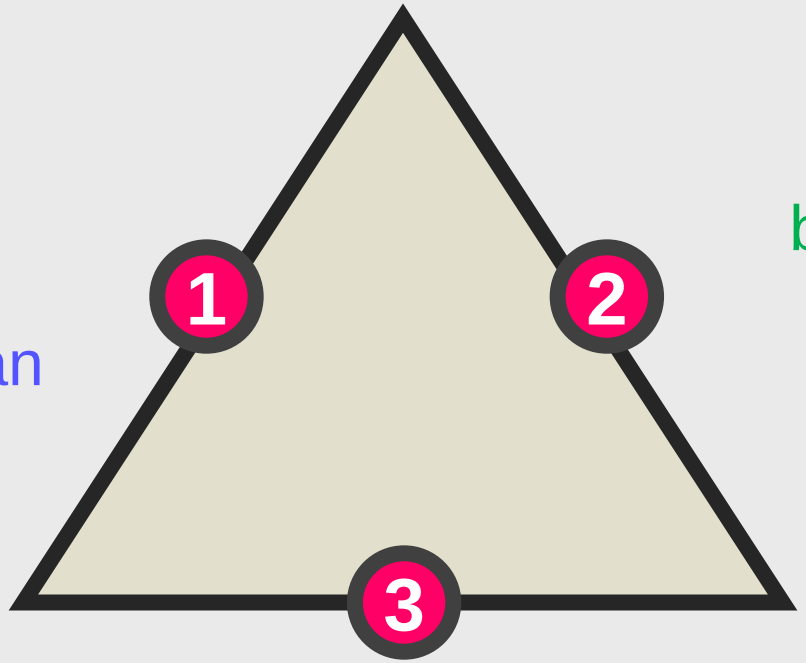
Rich and Knight, 1991

Dimensions of AI Definitions

Building intelligent artefacts

vs.

understanding human behaviour.



Should the system

behave like a human

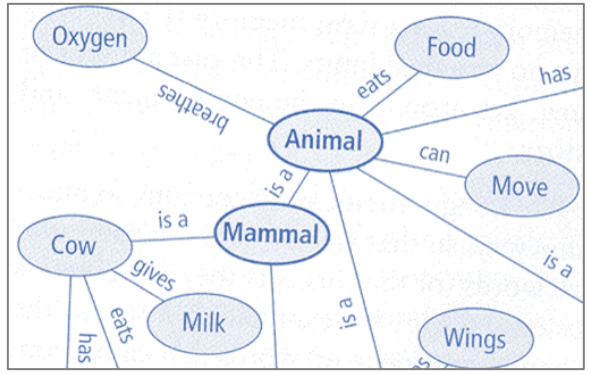
Or

behave *intelligently*?

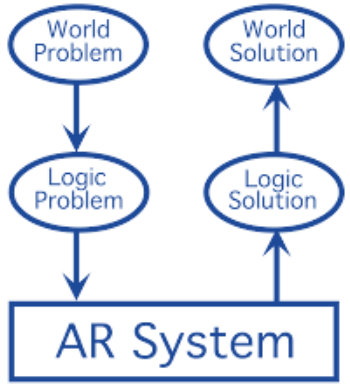
Does it matter how I built it

as long as it does the job well?

What Does AI Really Do?



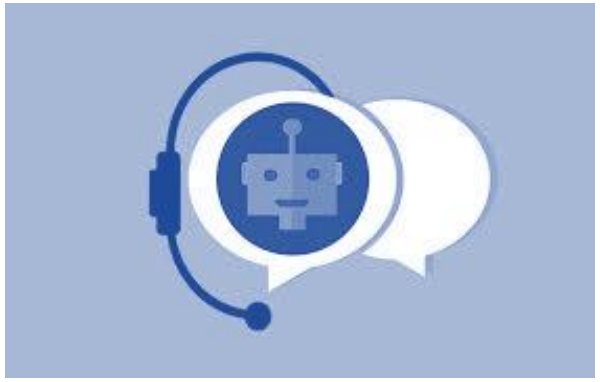
Knowledge Representation



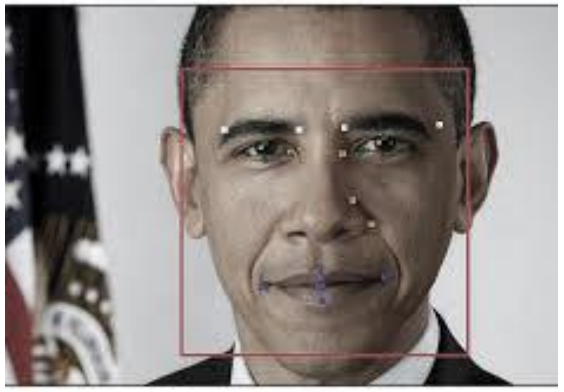
Automated reasoning

Planning

Machine Learning



Natural language understanding



Machine vision

Robotics

Google

Web Search

Alan Turing - Father of AI

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

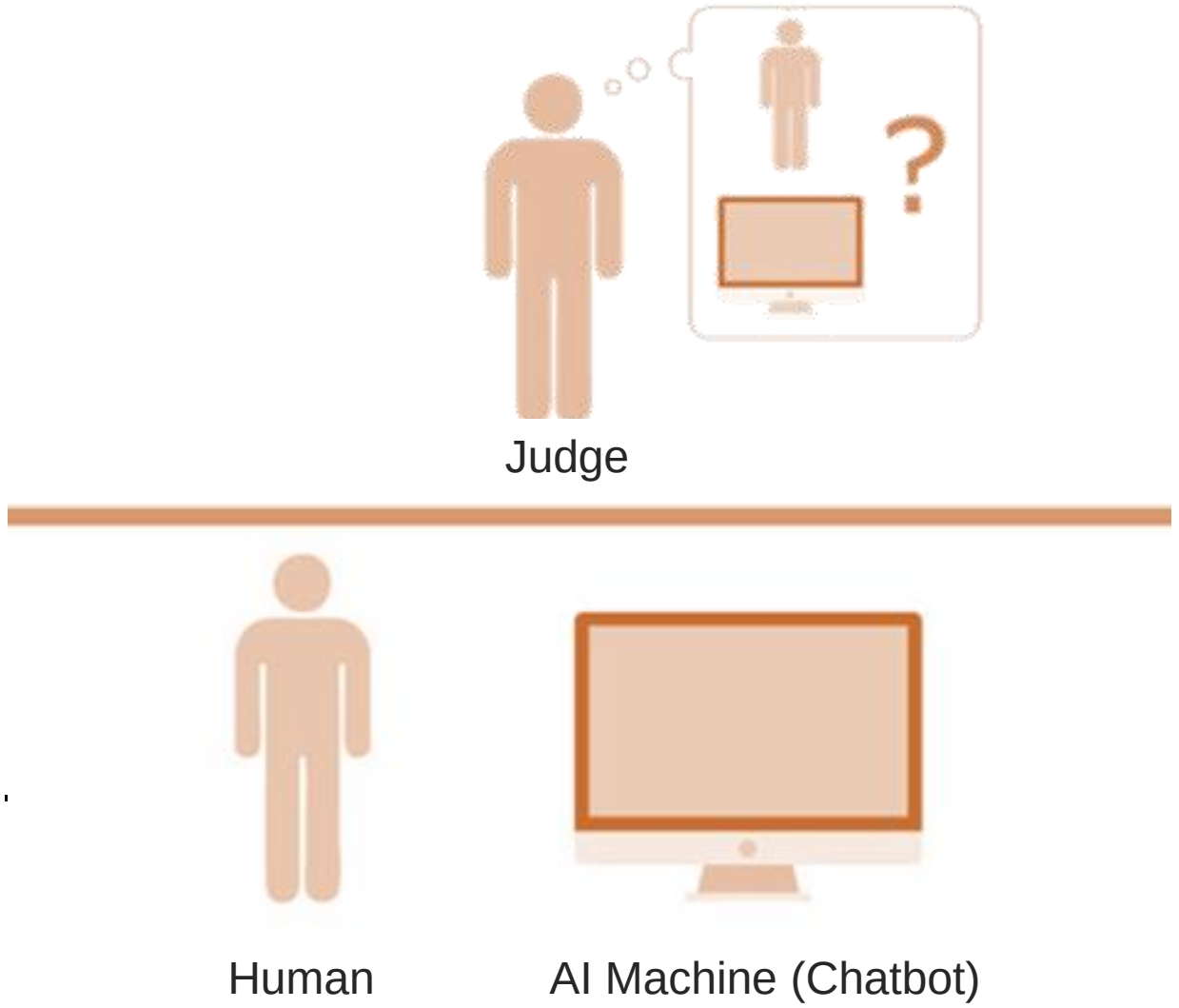
By A. M. Turing

I propose to consider the question, 'Can machines think?' ...

Turing, A.M. (1950), Computing machinery and intelligence, Mind, Vol.59, pp. 433-460

Turing Test

1. Judge (Human) communicates with a human and a machine over text-only channel.
2. Both human and machine try to act like a human.
3. Judge tries to tell which is which.



AI Definition Revisited

Systems that think like humans	Systems that think rationally
Systems that act like humans	**Systems that act rationally**

- Focus on action (act rationally).
- Avoids philosophical issues such as “is the system conscious.”
- Distinction may not be that important
  - acting rationally / like a human presumably requires (some sort of) thinking rationally / like a human,
  - humans much more rational in complex domains

Lessons from AI Research

What's Easy?

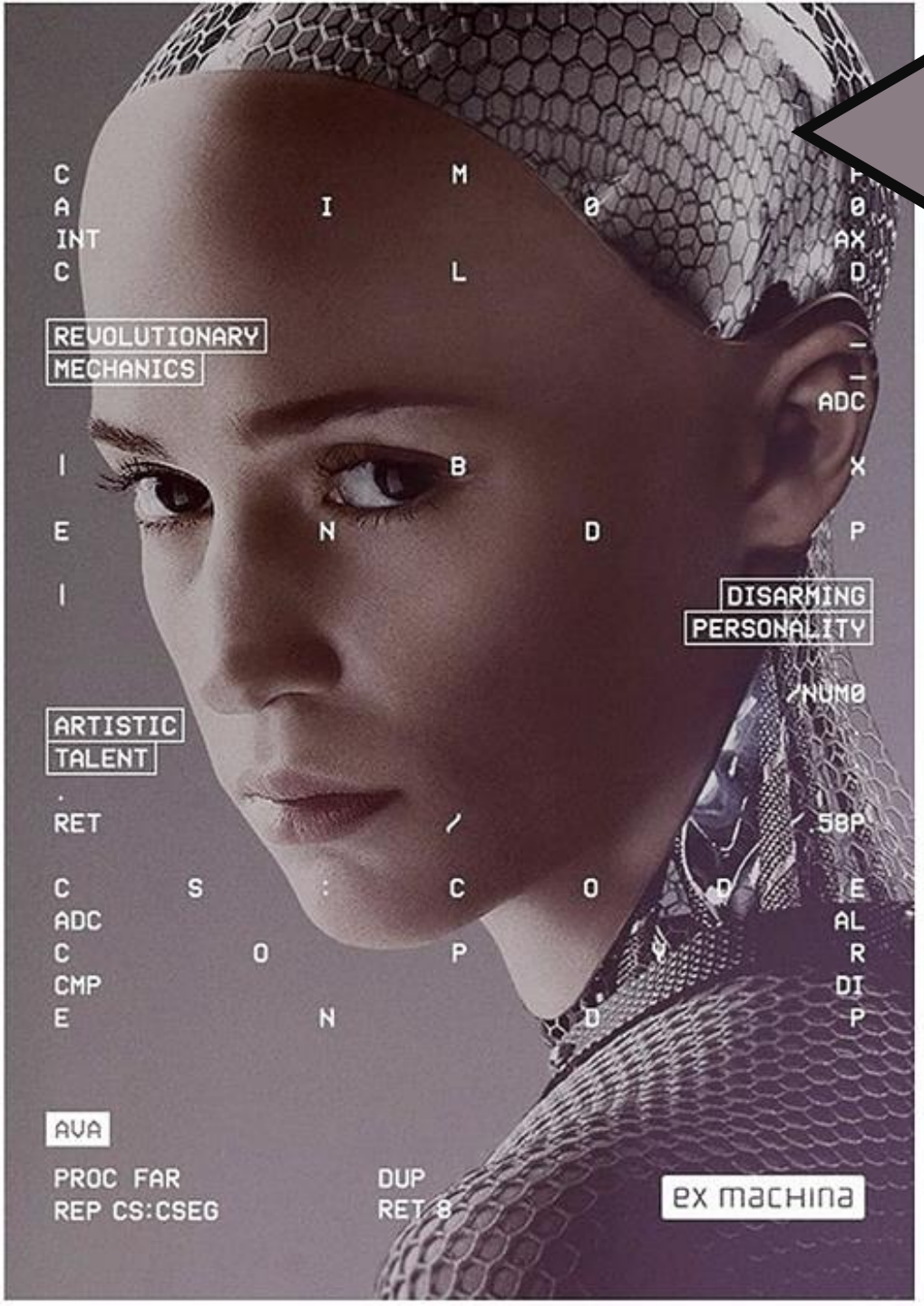
Clearly-defined tasks that we think require intelligence and education from humans tend to be doable for AI techniques

What’s Hard?

Complex, messy, ambiguous tasks that are natural for humans (in some cases other animals) are much harder

Types of AI

- **General-purpose AI** like the robots of science fiction is incredibly hard.
  - Human brain appears to have lots of special and general functions, integrated in some amazing way that we really do not understand at all (yet)
- **Special-purpose AI** is more doable (nontrivial?)
  - E.g., chess/poker playing programs, logistics planning, automated translation, voice recognition, web search, data mining, medical diagnosis, keeping a car on the road



The Goal

But busy in…

Puppy

or

muffin?

What Humans are Better At?

1. **Humans better at coming up with reasonably good solutions** in complex environments

2. **Humans better at adapting/self-evaluation/creativity** (“My usual strategy for chess is getting me into trouble against this person… Why? What else can I do?”)

Human		AI
Evolved for survival		Designed by engineers
Sets own goals		Goals programmed explicitly (usually)
Complex, general purpose system	v/s	Specific, constrained system
Continually learns		Can turn off learning, or not use learning
Learns from all observed data		Data access can be controlled
Learns only from own experiences		Can share data with other robots
Can make any choice at any time		Available actions can be restricted

Some AI Applications

https://www.youtube.com/watch?v=8IO6ED0p1Sk

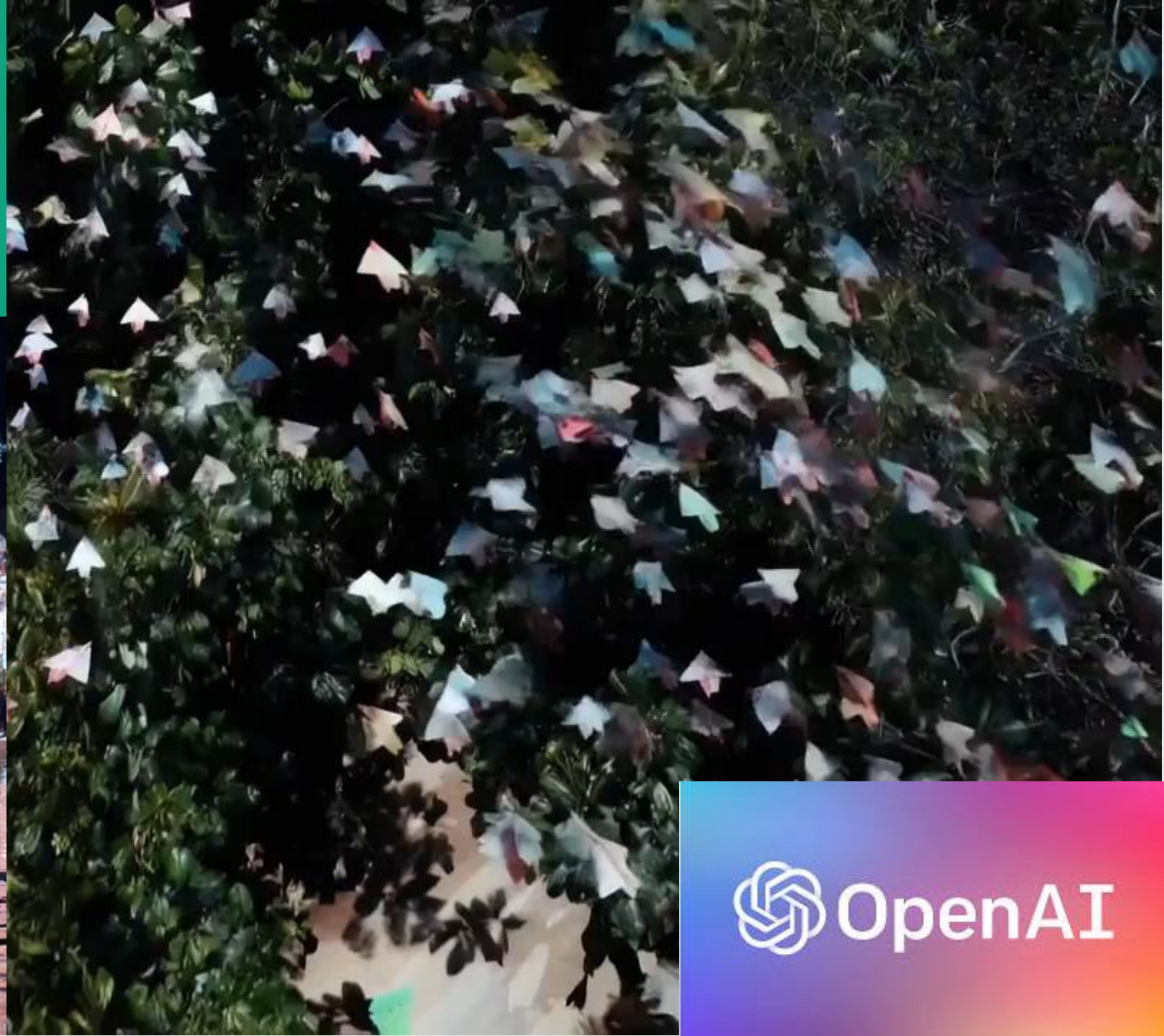
- Robotics
- Planning
- Navigation
- Search
- Optimisation
- Learning

Example AI Applications

- **Search**
  - Solving a Rubik’s cube
- **Constraint satisfaction/optimization problems**
  - Scheduling a given set of meetings (optimally)
- **Game playing**
  - Playing chess
- **Logic, knowledge representation**
  - Solving logic puzzles, proving theorems
- **Planning**
  - Finding a schedule that will allow you to graduate (reasoning backwards from the goal)
- **Probability, decision theory, reasoning under uncertainty**
  - Given some symptoms, what is the probability that a patient has a particular condition? How should we treat the patient?
- **Machine learning, reinforcement learning**
  - Recognizing handwritten digits

ChatGPT

https://openai.com/index/sora/

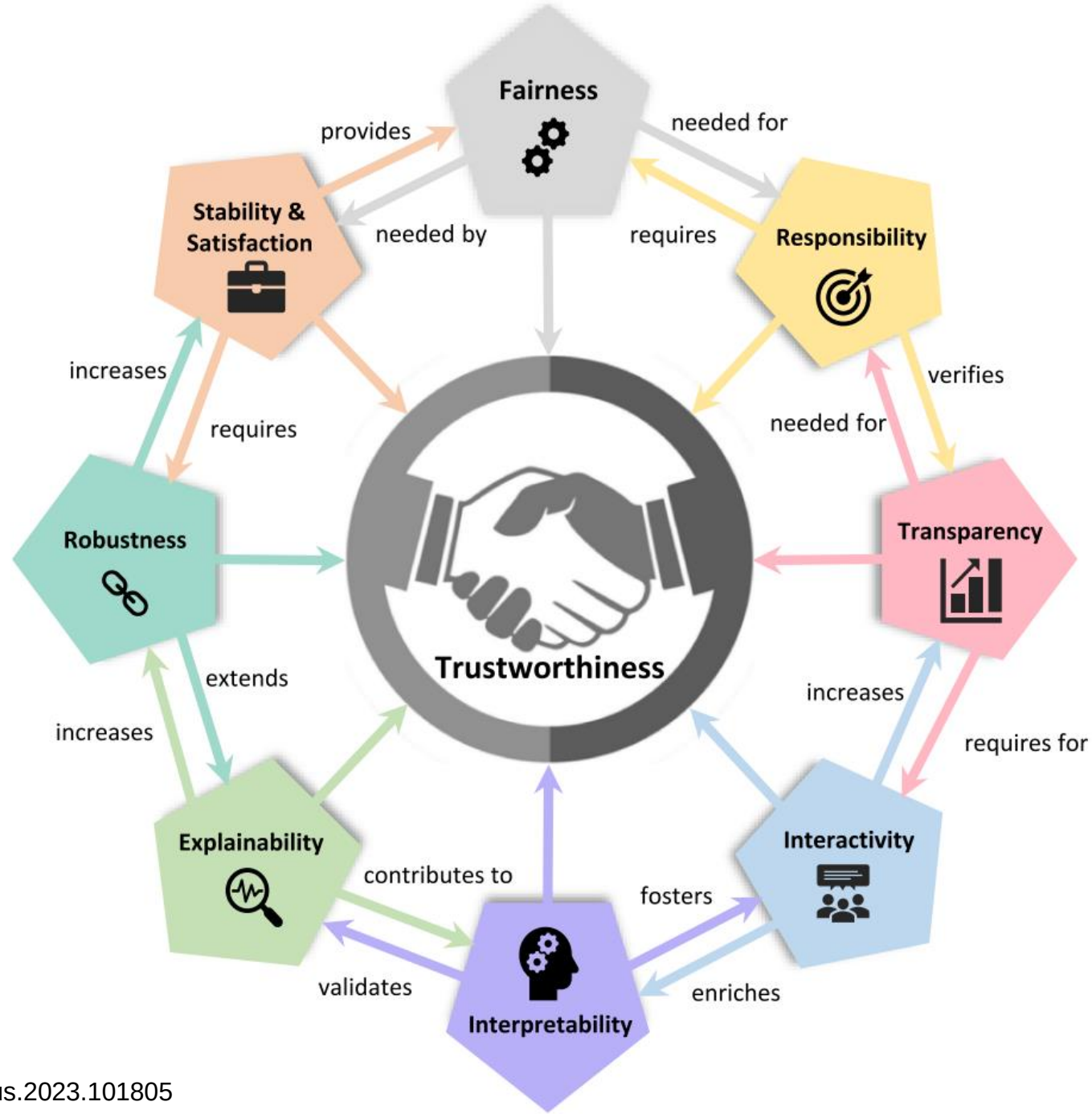


OpenAI

The trolley problem

THE CIRCLE

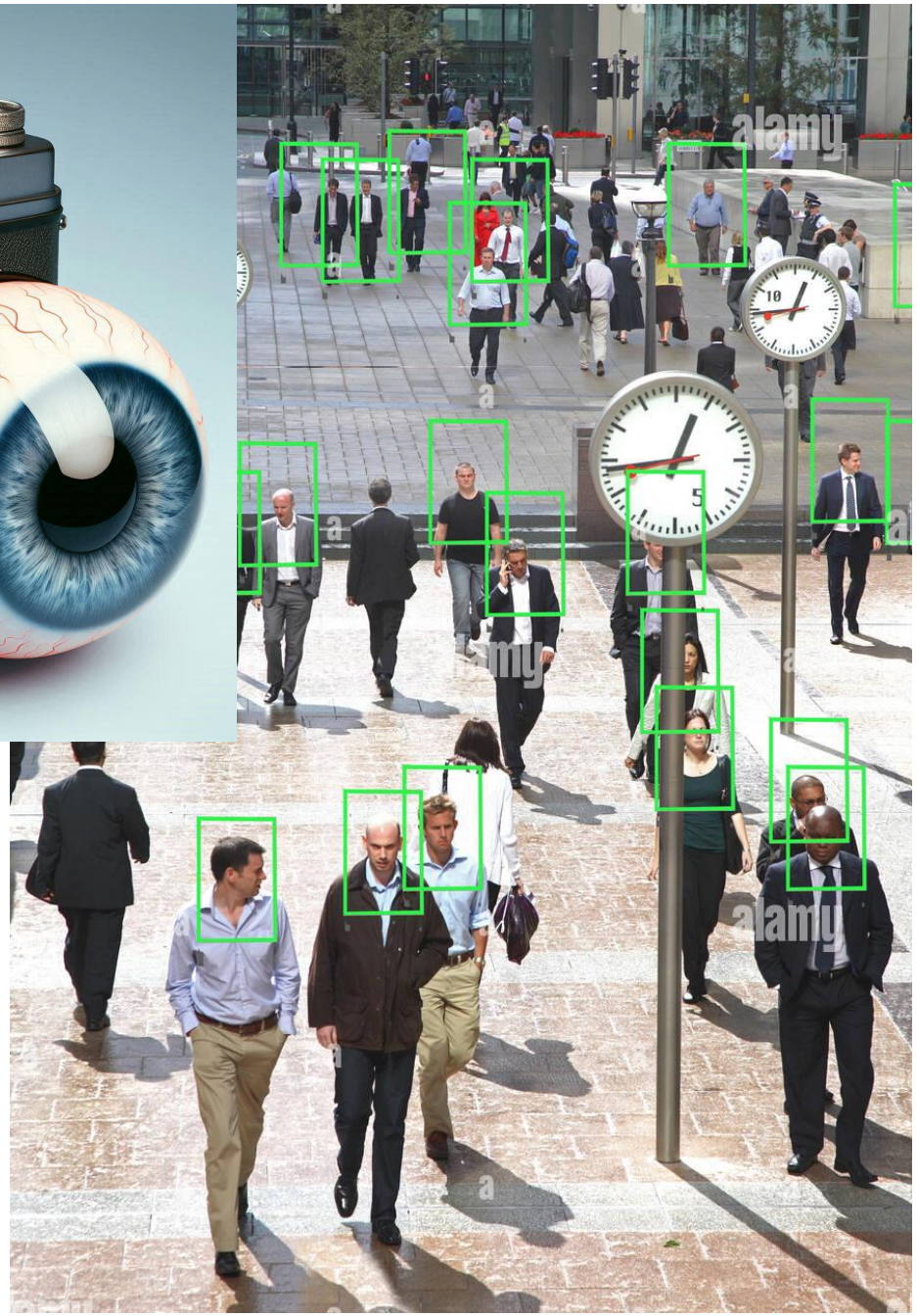
Knowing is Good
Knowing Everything is Better



Source: https://doi.org/10.1016/j.inffus.2023.101805


by DALL·E 3

Privacy



my phone when i say i want to buy something:

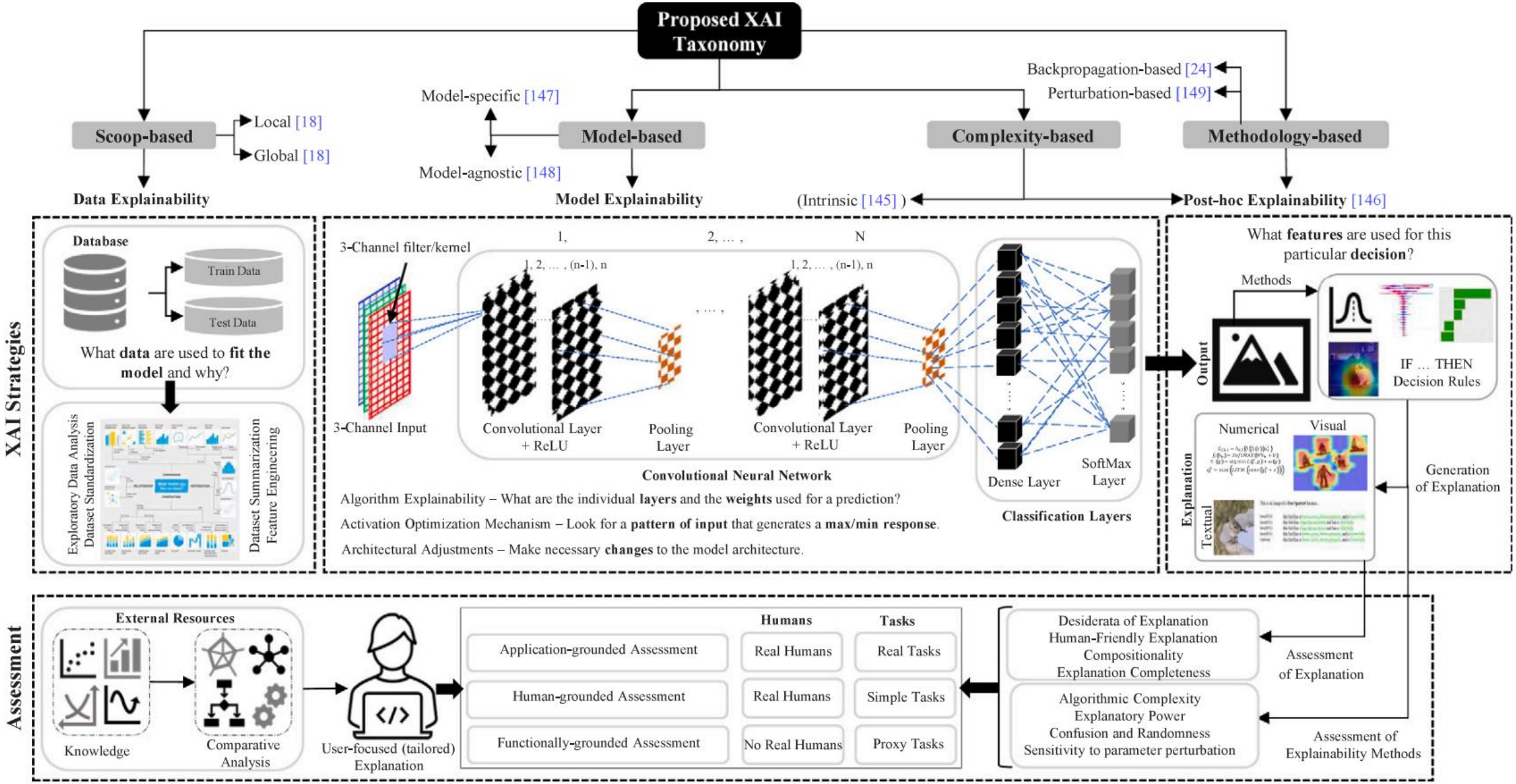
Bias

Classification Accuracy.
The worst recognition was on darker females, failing on over 1 in 3 women of colour.

A key factor in the accuracy differences is the lack of diversity in training images and benchmark data sets.

(source: https://physicsworld.com/a/fighting-algorithmic-bias-in-artificial-intelligence/)

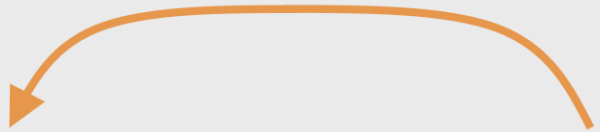
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Source: https://doi.org/10.1016/j.inffus.2023.101805

Post-model explanations

Holding AI developers accountable requires an understanding of how their AI systems work.

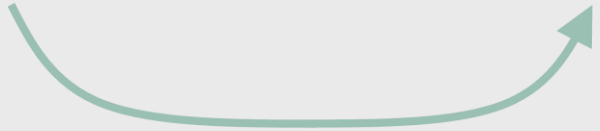


Accountability

Responsibility for one's actions and their consequences, whether positive or negative

Transparency

The act of being open, often via clear and honest communication



The necessary level of transparency is informed by the accountability objective.

AlphaGo

AlphaGo mastered the ancient game of Go, defeated a Go world champion, and inspired a new era of AI systems.

https://deepmind.google/technologies/alphago/

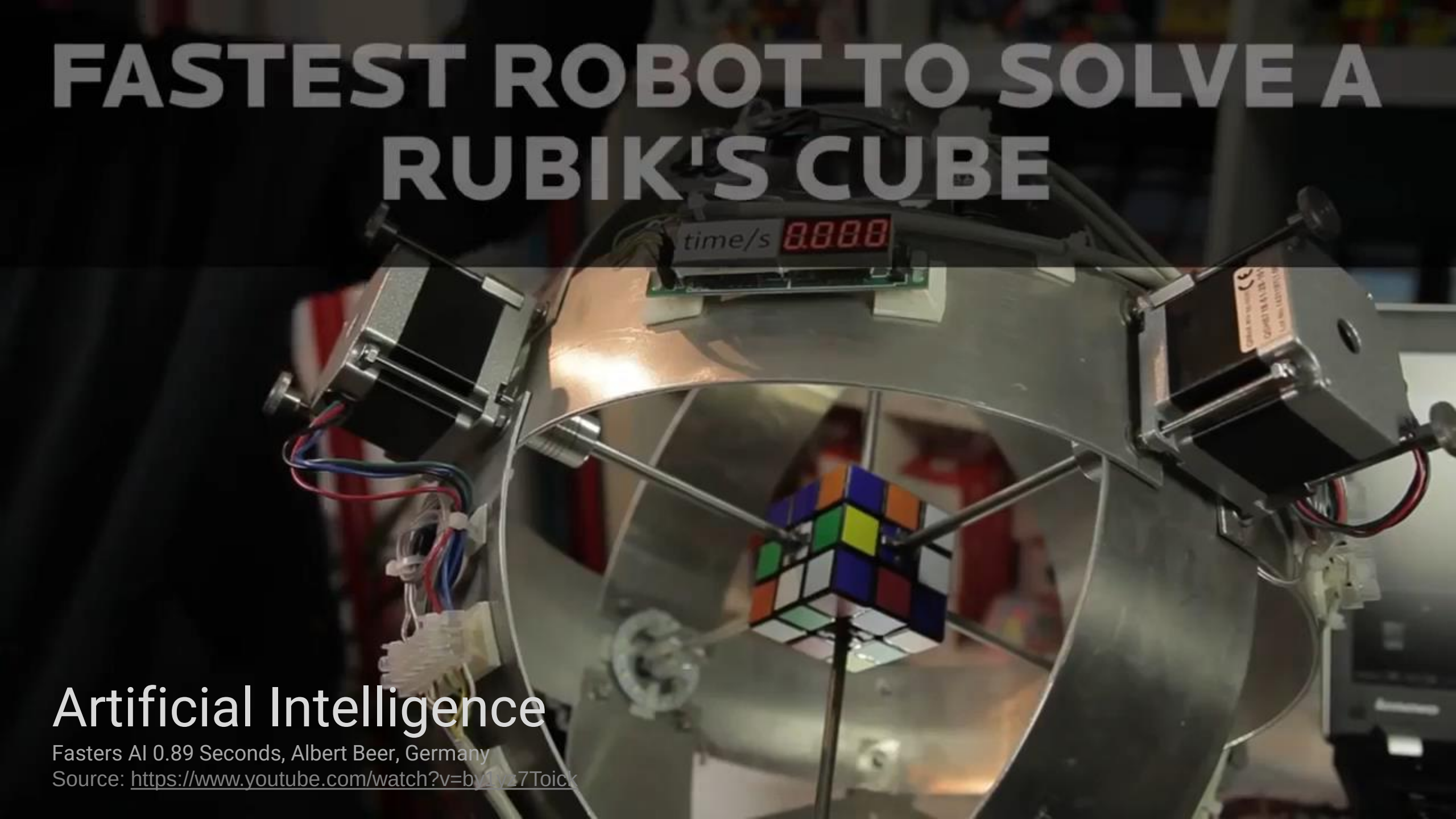
Human Vs AI

Rubik’s Cube is a search problem

Human Intelligence

Rubik's Cube World Record 4.73 Feliks Zemdegs

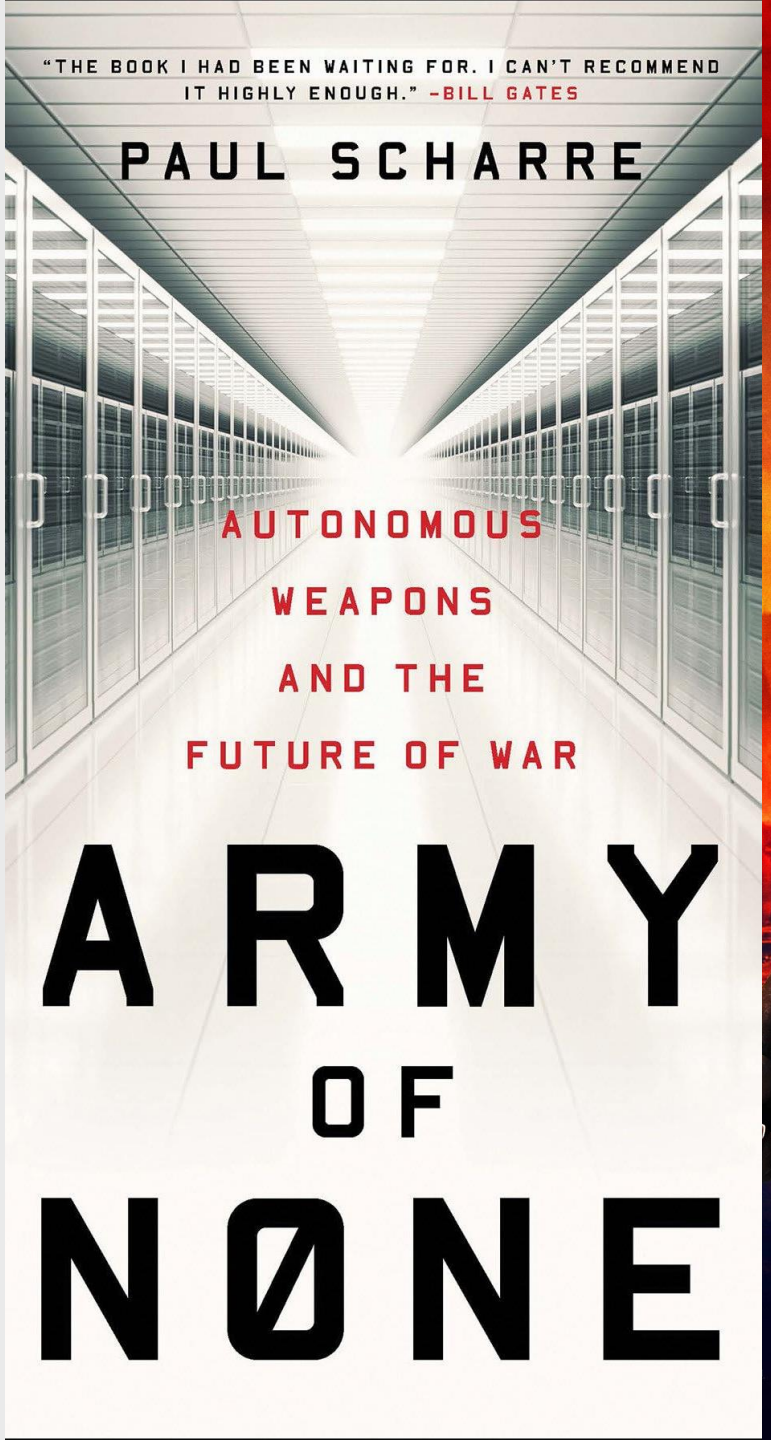
Source: https://www.youtube.com/watch?v=M5yjkpMXChI



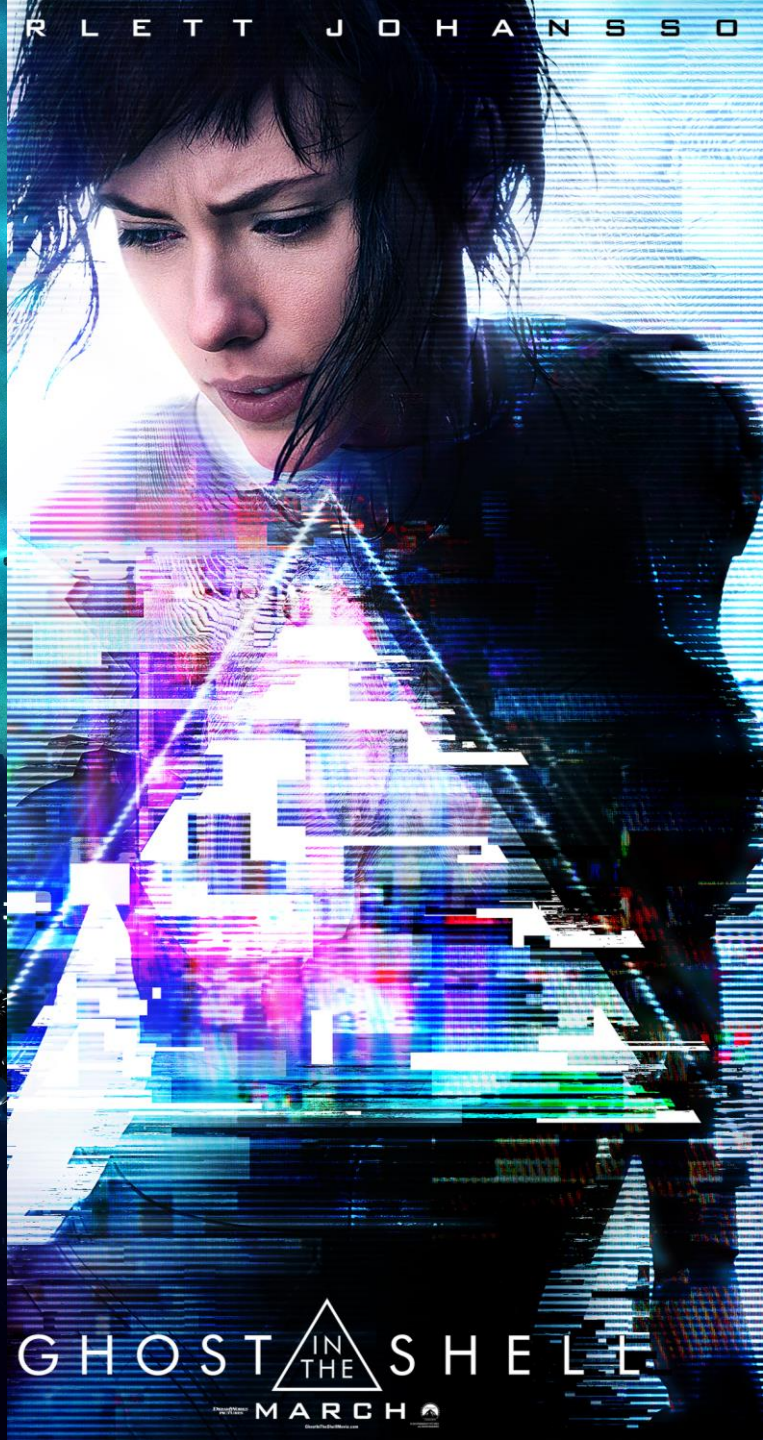
Artificial Intelligence

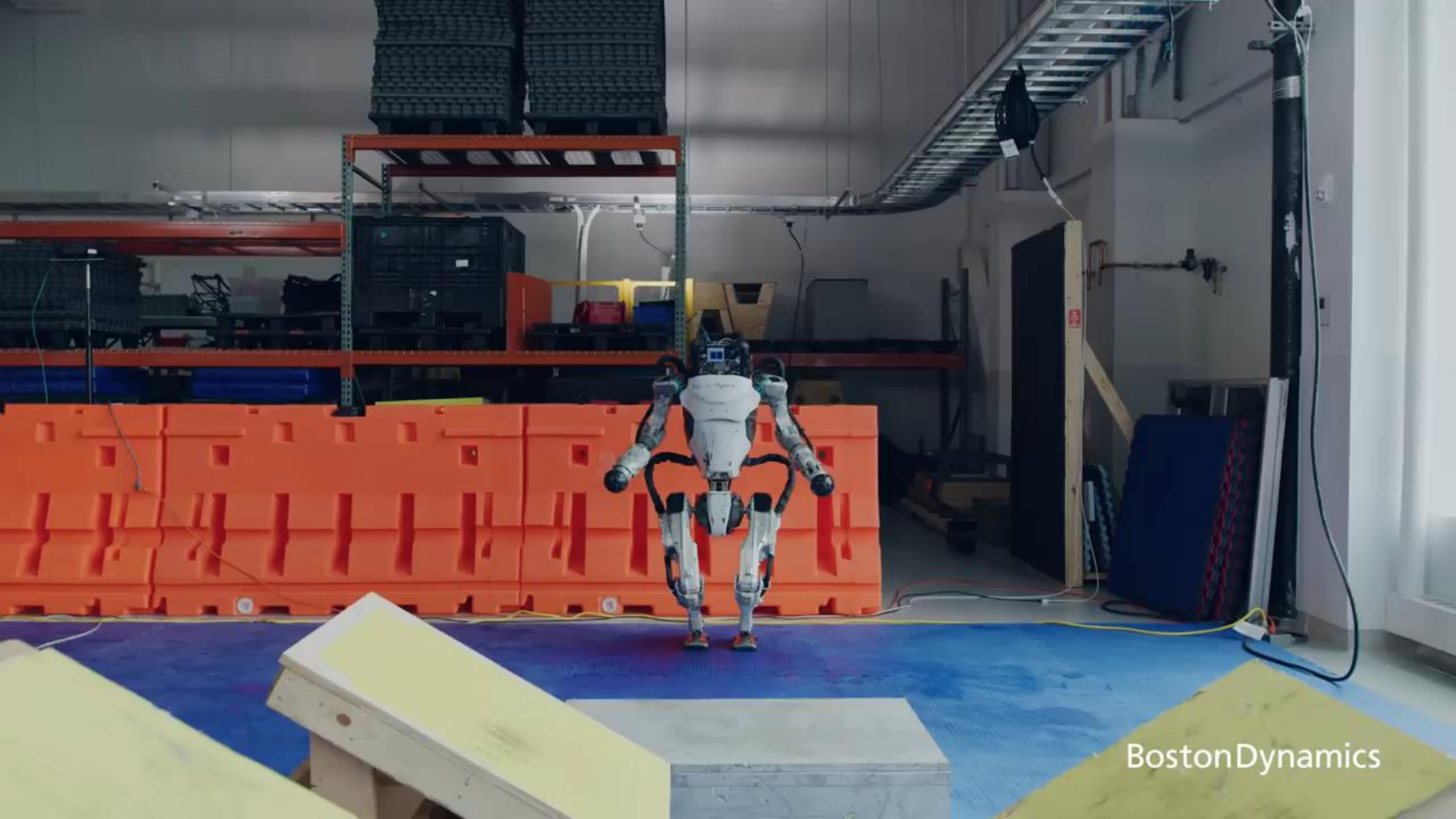
Fasters AI 0.89 Seconds, Albert Beer, Germany
Source: https://www.youtube.com/watch?v=by1yz7Toick

Risk of AI?


"THE BOOK I HAD BEEN WAITING FOR. I CAN'T RECOMMEND IT HIGHLY ENOUGH." -BILL GATES
PAUL SCHARRE
AUTONOMOUS WEAPONS AND THE FUTURE OF WAR
ARMY OF NØNE


RYAN GOSLING HARRISON FORD
BLADE RUNNER 2049
EXPERIENCE IT IN IMAX


GHOST IN THE SHELL
MARCH


Boston Dynamics


Clear Documentation
Ownership of failure
Explainability
Accountability in AI
Auditing and Testing
External Review
Data Governance

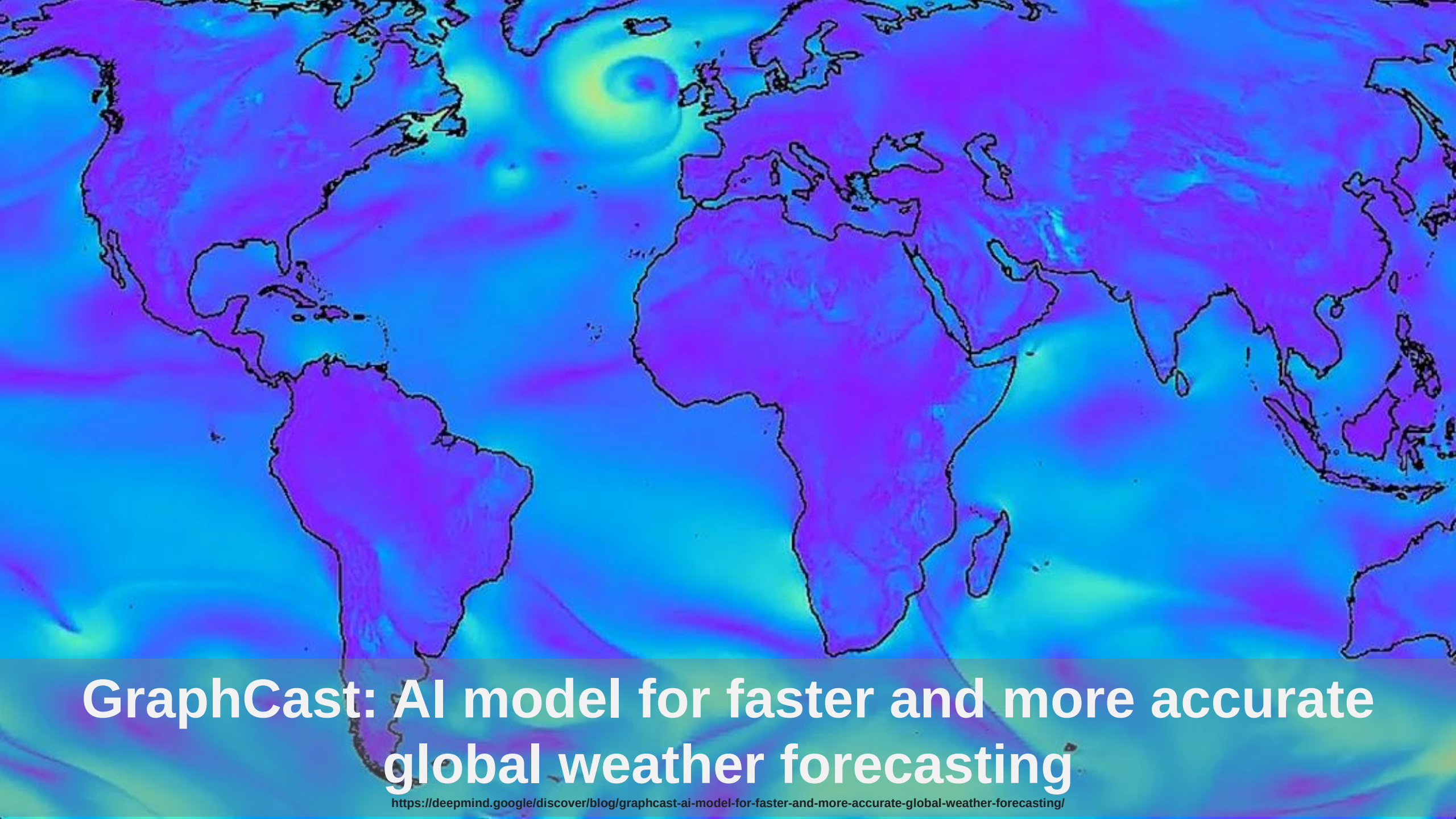
Usefull Applications of AI

Benefits of AI to the Society

AlphaFold

The Protein Folding Problem:
A failure in protein folding causes several known diseases, and scientists hypothesize that many more diseases may be related to folding problems.

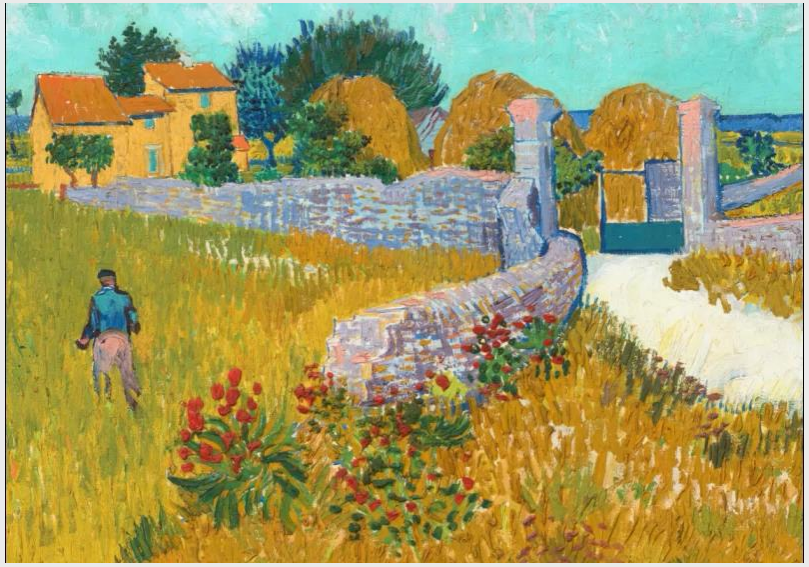
https://deepmind.google/technologies/alphafold/

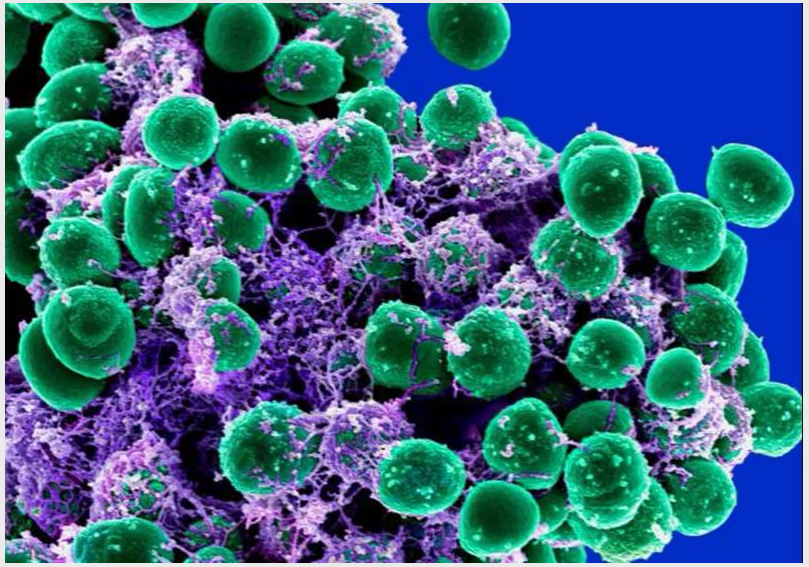


GraphCast: AI model for faster and more accurate global weather forecasting

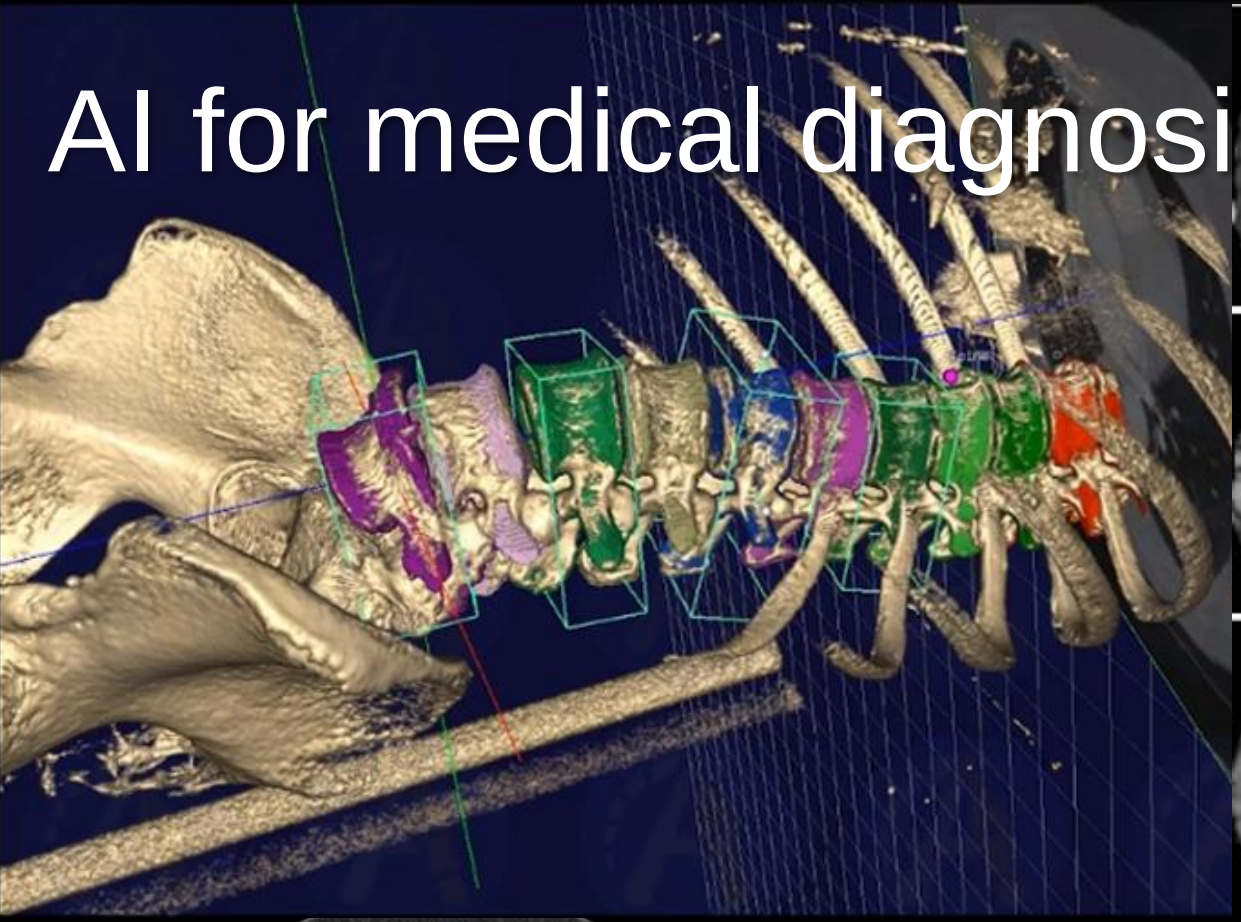
https://deepmind.google/discover/blog/graphcast-ai-model-for-faster-and-more-accurate-global-weather-forecasting/

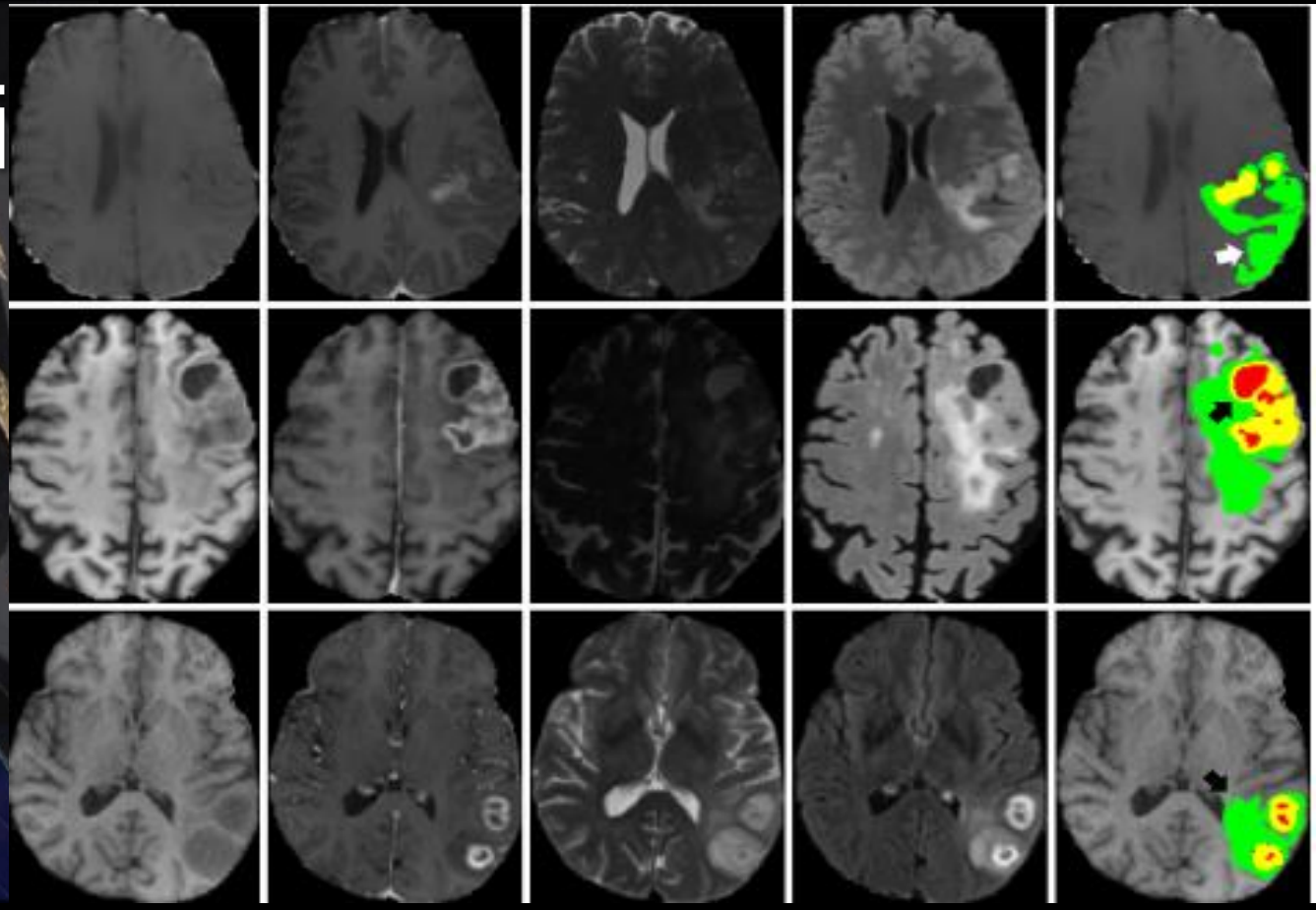
Deep Leaning Models:
Segment Anything from Meta (2023)





AI for medical diagnosi





AI Revolution in Brain Tumor Analysis: Harnessing cross-spectral multimodal inputs for advanced 3D semantic segmentation.

Precision in Tumor Detection: Enhancing 3D segmentation accuracy with our innovative AI-driven multimodal approach.

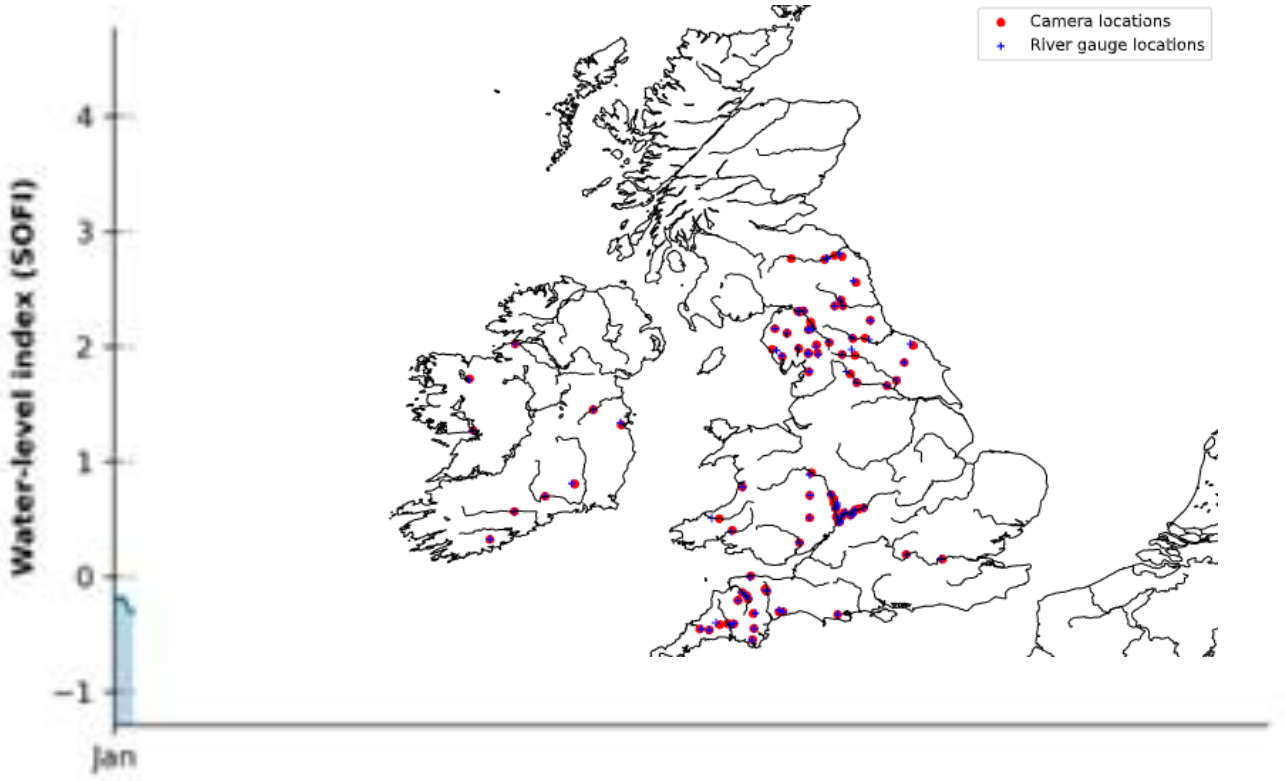
Futuristic Neuroimaging: Elevating 3D brain tumor detection accuracy with our cutting-edge AI model.

Redefining Medical Imaging: Cross-spectral multimodal inputs setting new standards in 3D brain tumor semantic segmentation.

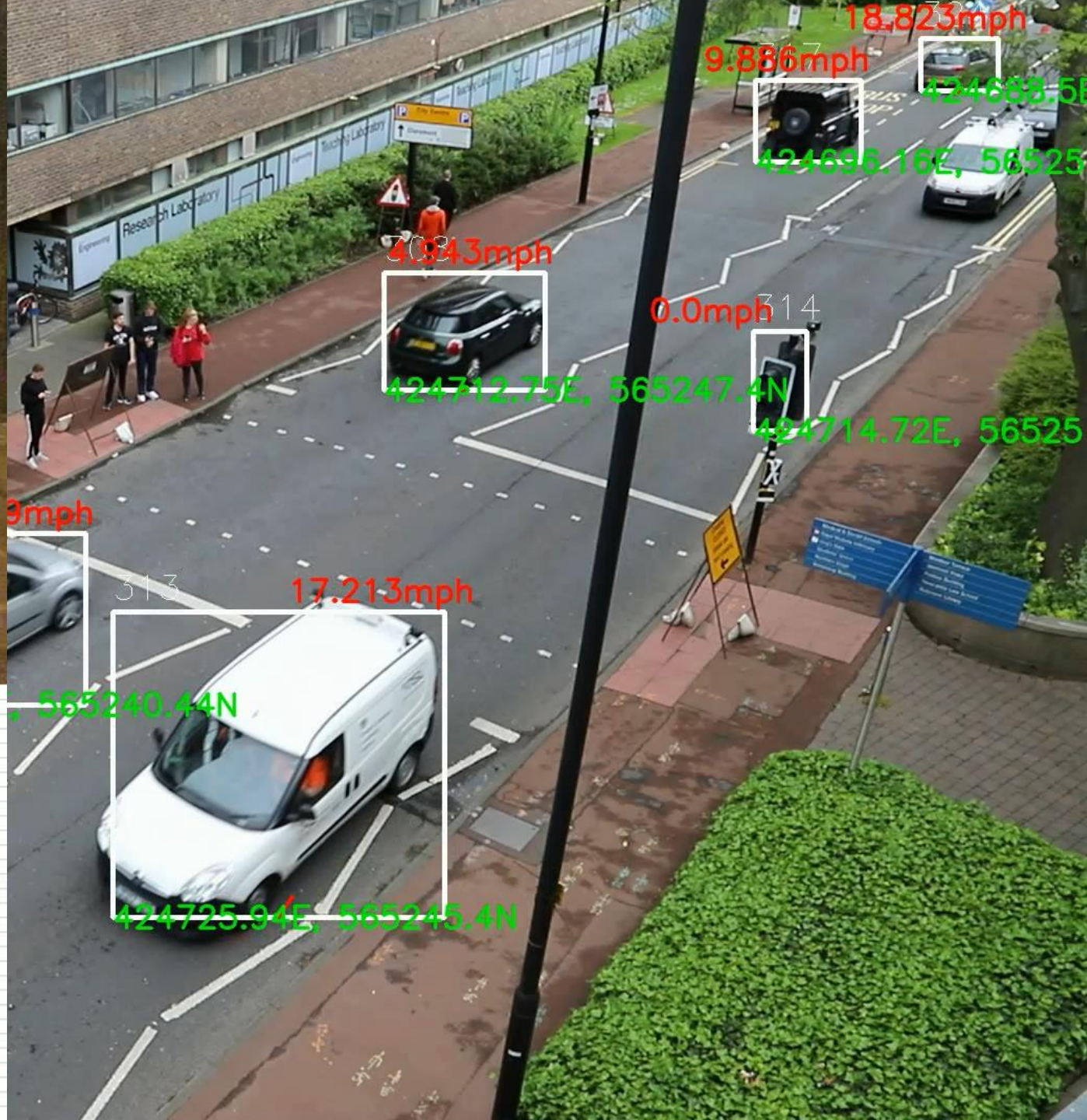
Clarity in Brain Health: Unveiling a new era in medical imaging for brain tumor segmentation with our AI solution

AI for Automated Flood Tracking

Vandaele, Dance, and Ojha, (2021) *Hydrology and Earth System Sciences*

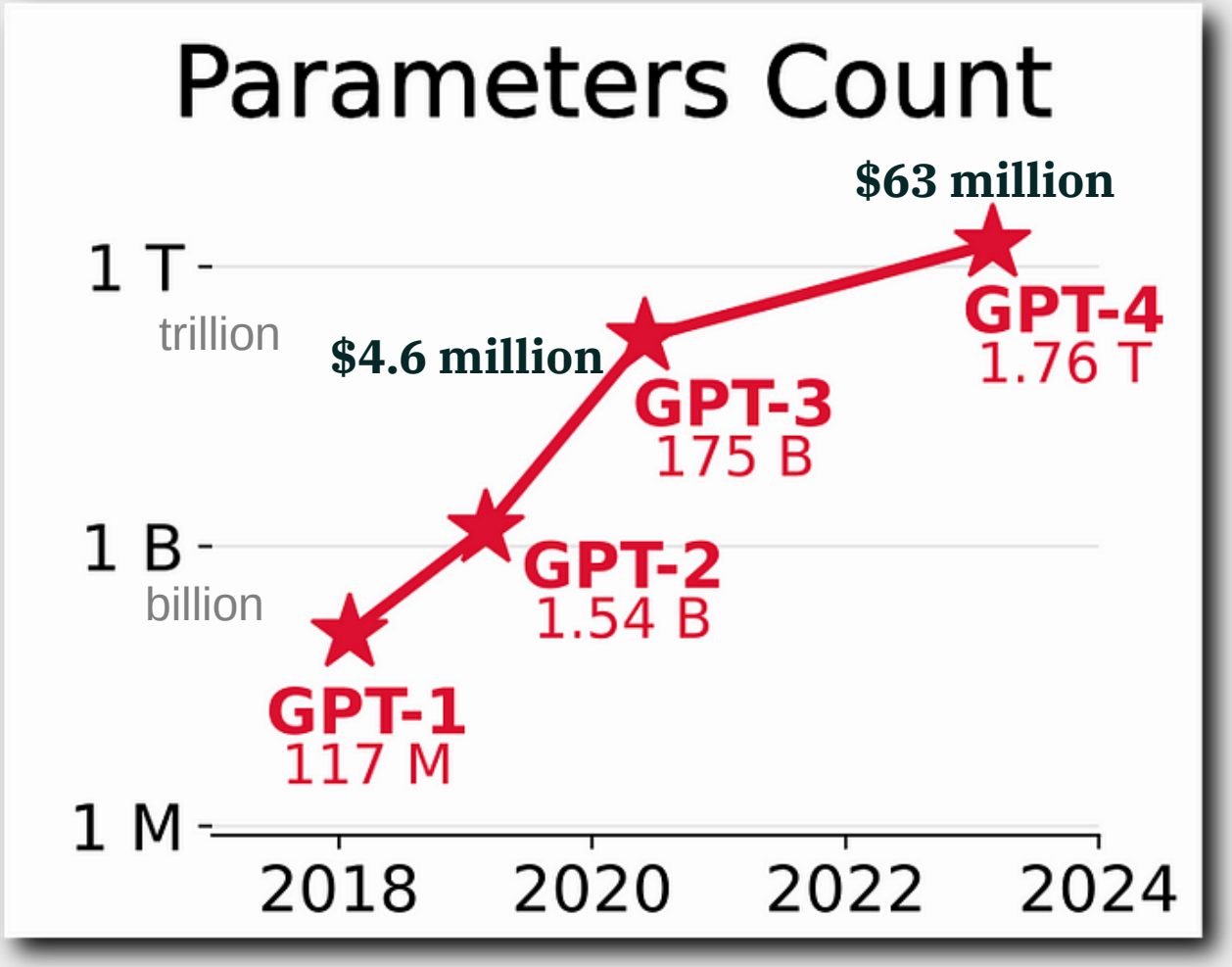


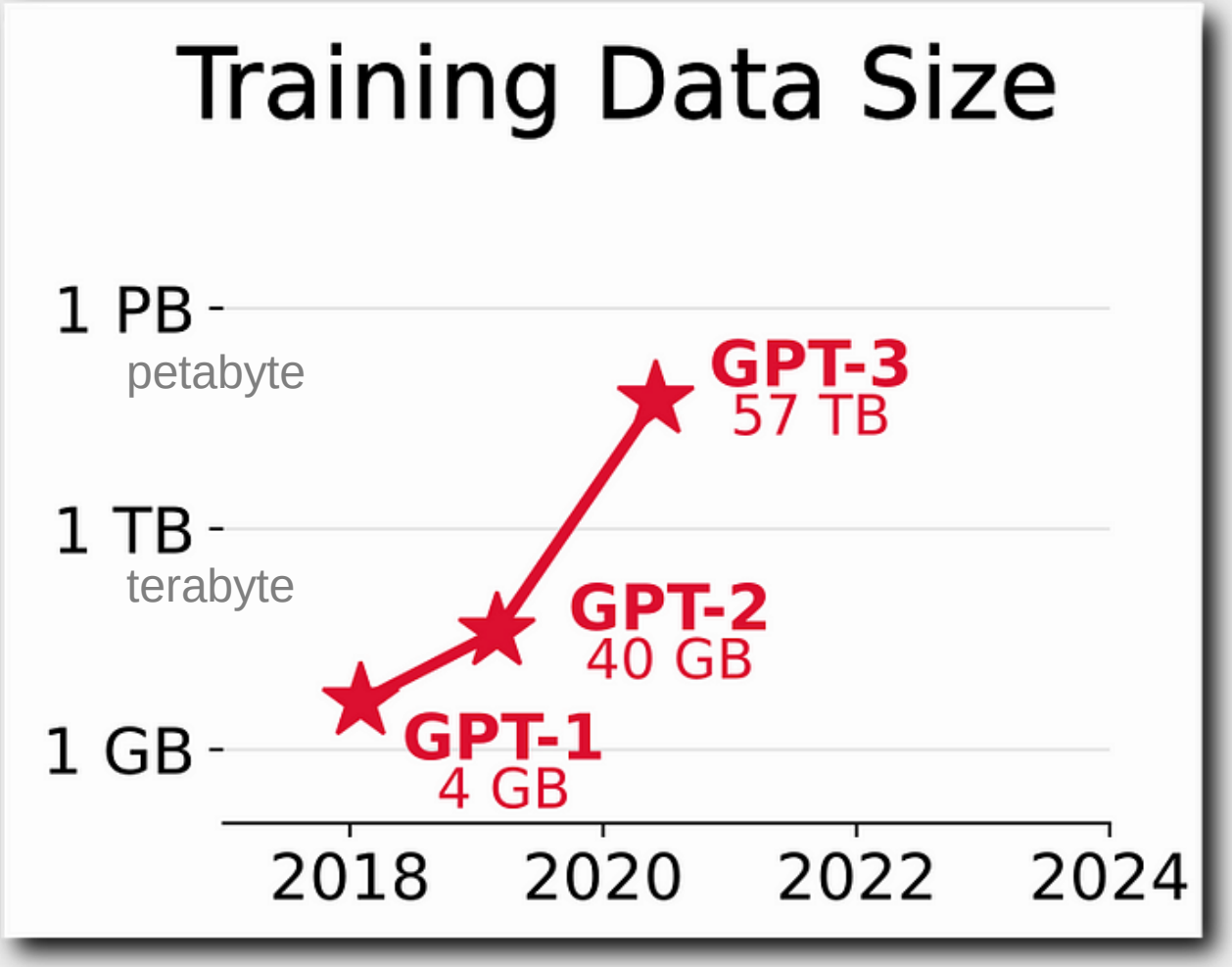
Smart City



Source: Phil et al (Newcastle)

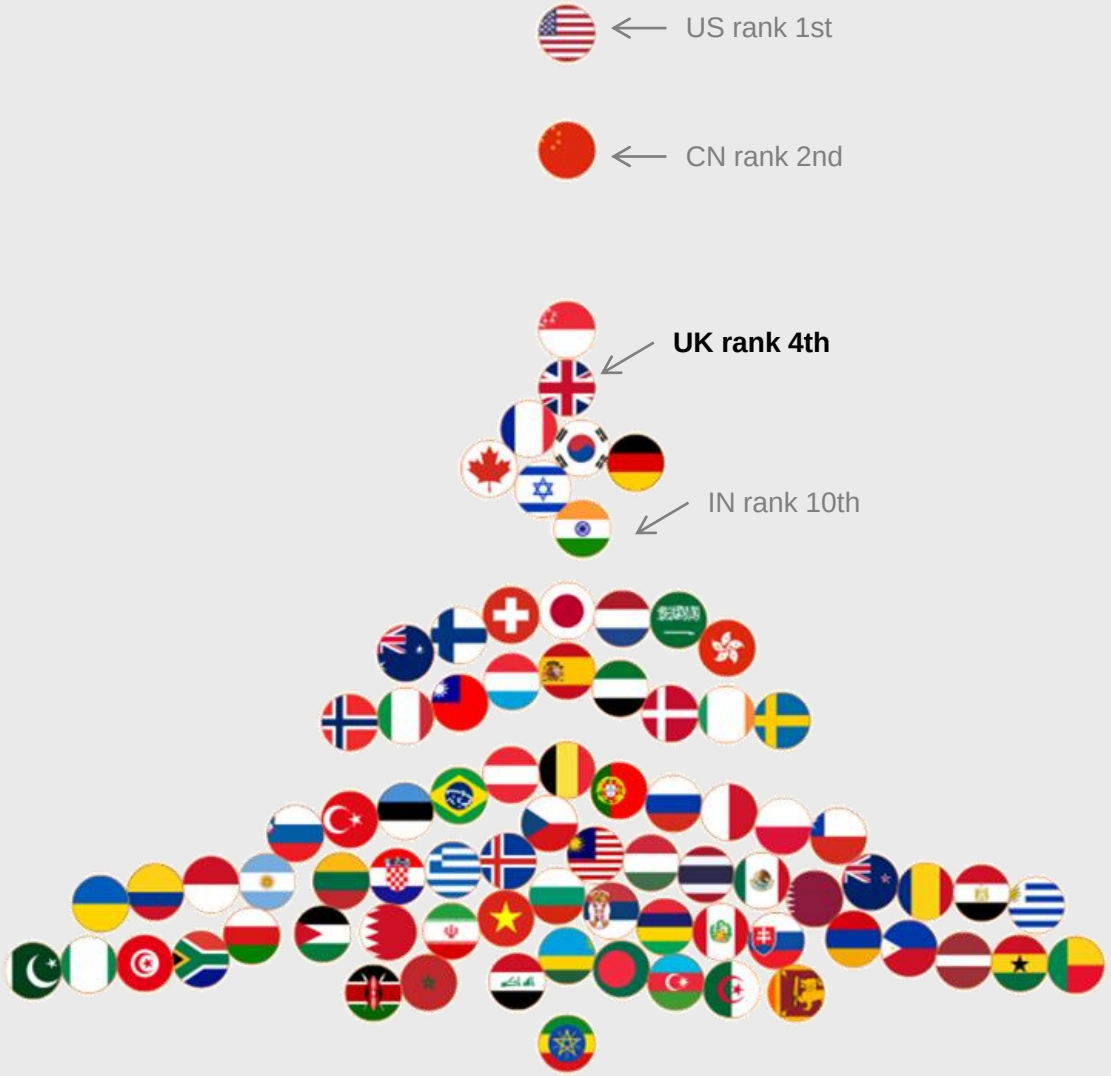
Massive AI Model and Data Size





Source: Francesco Casalegno, ChatGPT Unveiled: What's the ML Model Inside it, from GPT-1 to GPT-4

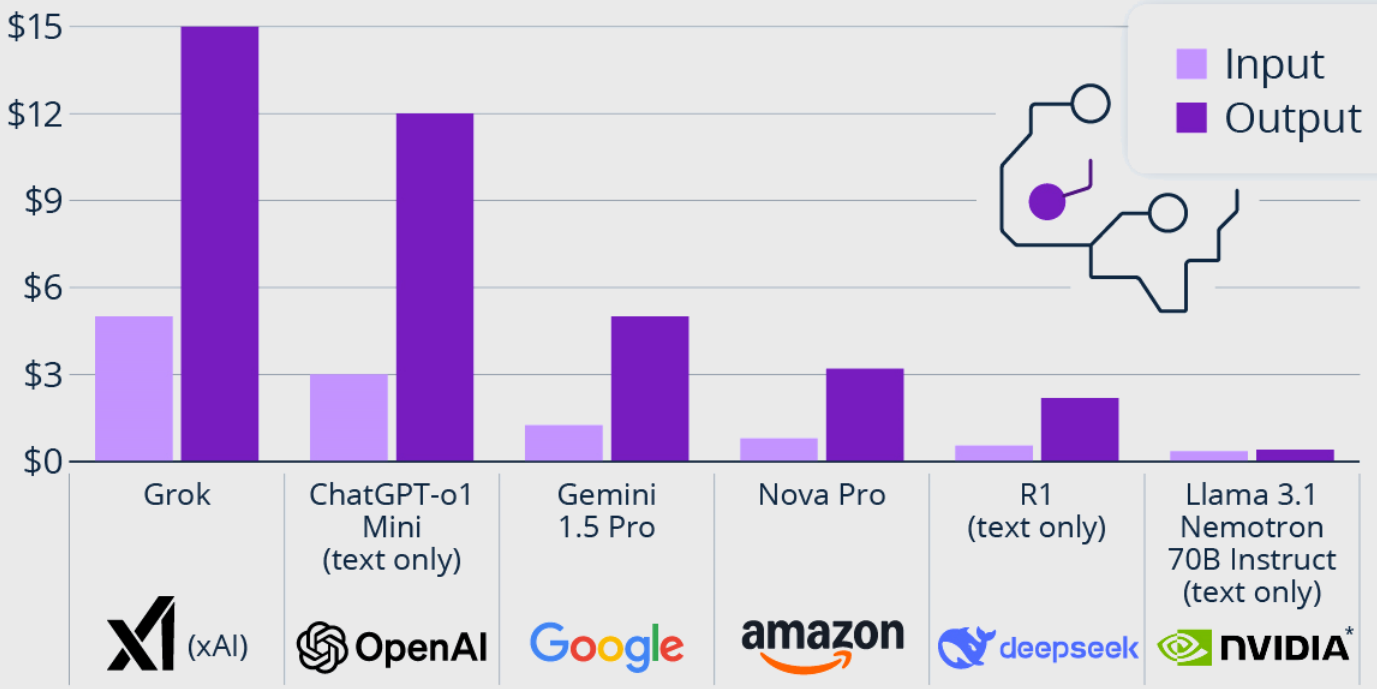
AI Arms Race



https://www.tortoisemedia.com/intelligence/global-ai

DeepSeek-R1 Upsets AI Market With Low Prices

Estimated price for processing one million input/output tokens on different AI models



A token is the smallest unit of AI model processing (~4 characters).
o1 is ChatGPT's latest model. List includes most comparable model per company
* Uses Meta's open-source Llama AI

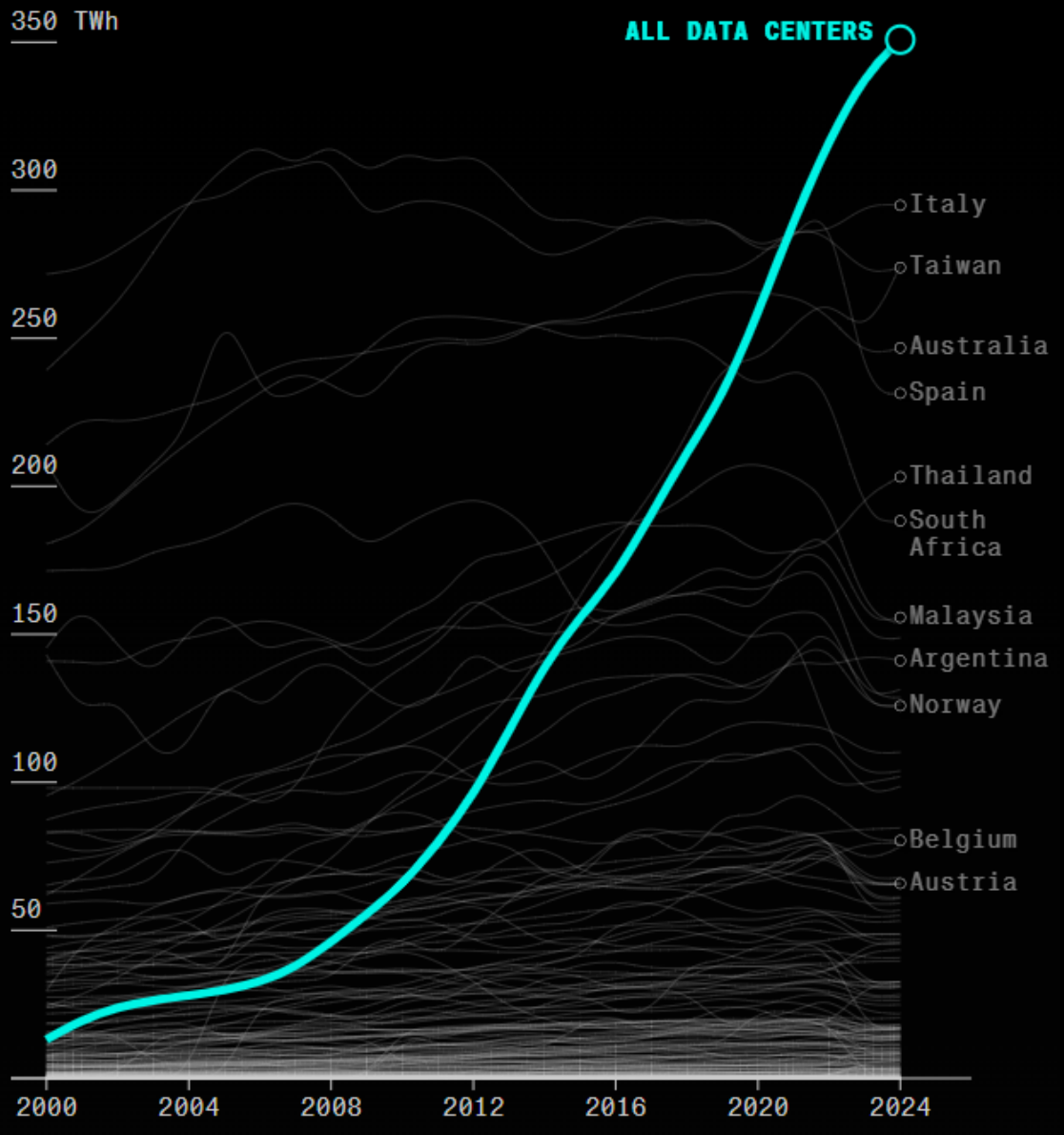
Source: DocsBot

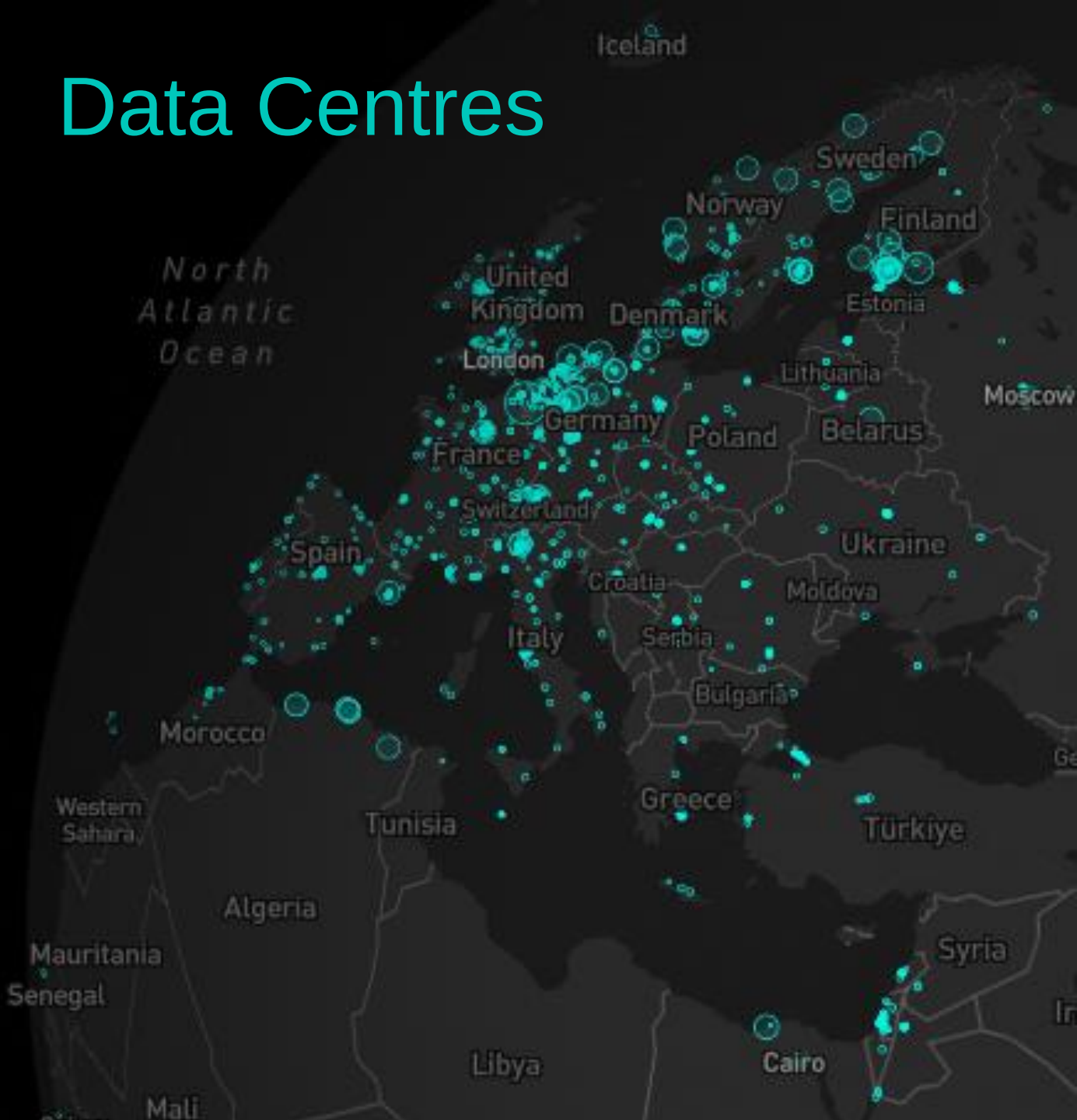
Data Centres

Altogether, data centers use more electricity than most countries

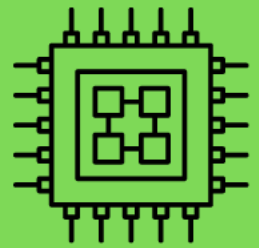
Only 16 nations, including the US and China, consume more

Source: Bloomberg





Elevating smaller models



Alternate deployment strategies

Carbon-efficiency and carbon-awareness

Sustainable AI

“It's going to be interesting to see how society deals with artificial intelligence, but it will definitely be cool.”

— Colin Angle

American businessperson, CEO iRobot