70 UNIVERSITÀ DEL SALENTO 1955-2025 UNI SALENTO

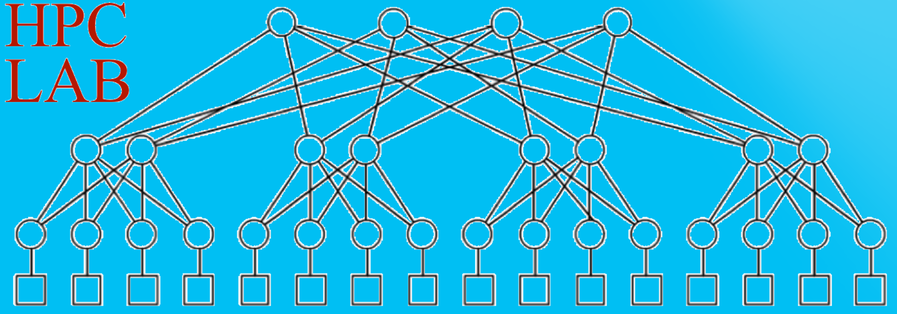
ARTIFICIAL INTELLIGENCE RISKS AND CHALLENGES

CSR INTERNATIONAL SUMMER SCHOOL

LECCE, JULY 15 2025

PROF. MASSIMO CAFARO, PH.D.

HPC LAB



OVERVIEW

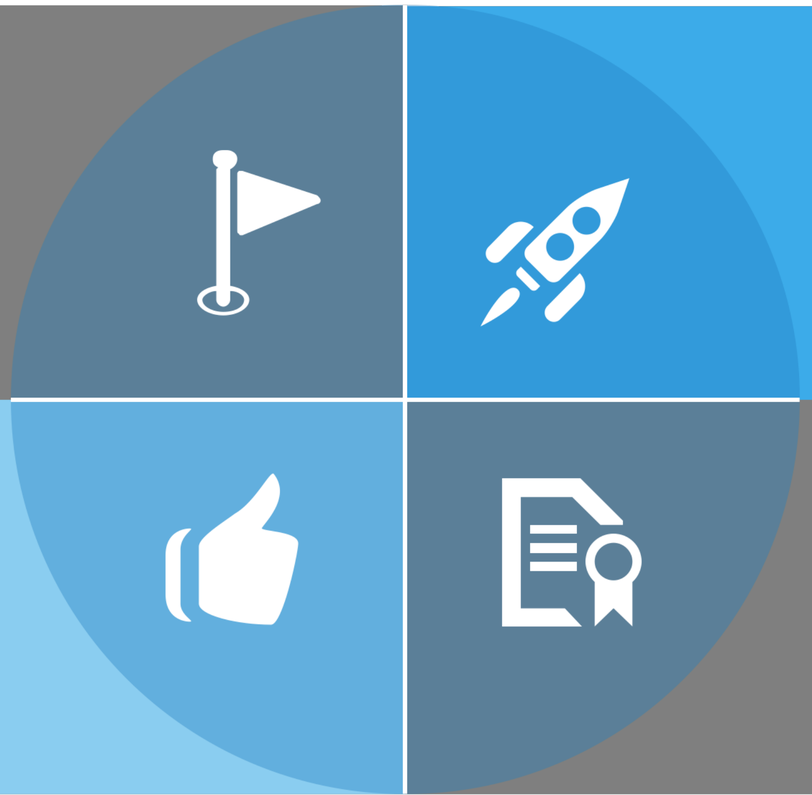
- What is AI
- People’s and scientists’ attitude
- Risks & challenges
- Conclusions

WHAT IS AI

WHAT AI REALLY IS

Over time, various definitions of Artificial Intelligence have been provided, as a science that studies and aspires to create machines that...

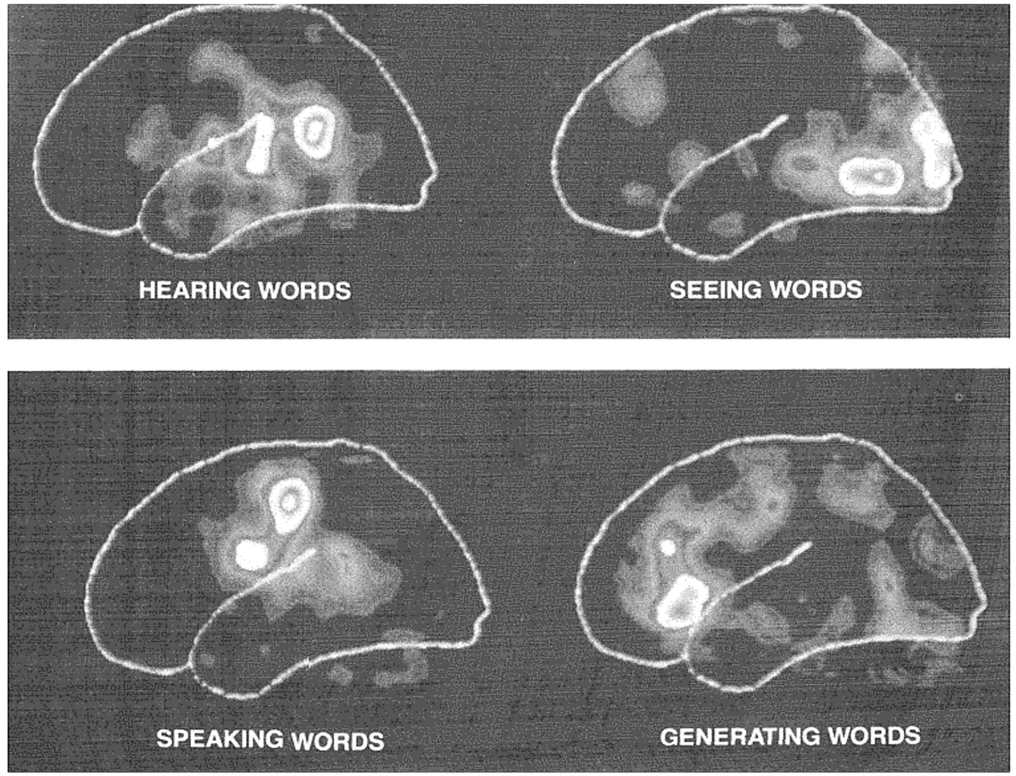
Think like humans	Think rationally
Act like humans	Act Rationally



THINKING LIKE HUMANS

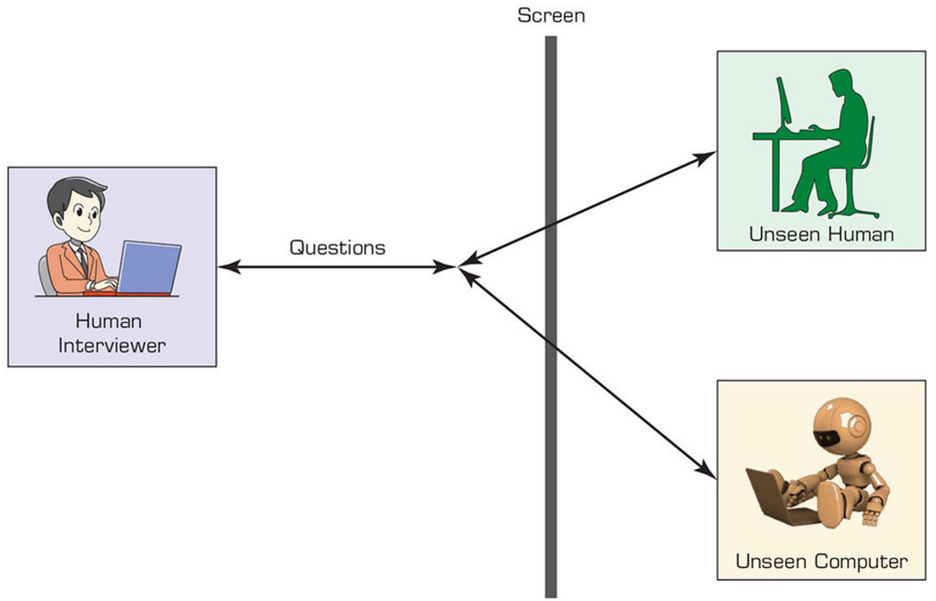
- The cognitive science approach
  - 1960s “cognitive revolution”: information-processing psychology replaced prevailing orthodoxy of behaviourism

- Scientific theories of internal activities of the brain
  - What level of abstraction? “Knowledge" or “circuits”?
  - Cognitive science: Predicting and testing behavior of human subjects (top-down)
  - Cognitive neuroscience: Direct identification from neurological data (bottom-up)
  - Both approaches now distinct from AI
  - Both share with AI the following characteristic: the available theories do not explain (or engender) anything resembling human-level general intelligence



ACTING LIKE HUMANS

- Alan Turing (1950) “Computing machinery and intelligence”
- “Can machines think? Can machines behave intelligently?”
- Operational test for intelligent behavior: the Imitation Game
  - If a machine could fool humans into thinking it’s a human, then it must be at least as intelligent as a normal human.
  - Turing predicted that by the year 2000 a program would be made which would fool the “average interrogator” 30% of the time after five minutes of questioning.
- To pass the Turing Test, a computer needs to be able:
  - to understand a human language (NLP)
  - to possess human intelligence (e.g., have a knowledge base)
  - to reason using its stored knowledge
  - to learn from its experiences (machine learning)



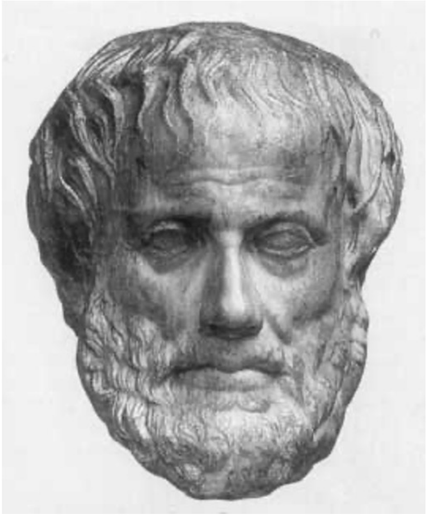
Measuring AI: the Turing test

The Leobner prize ($100,000 reward for the first program that judges cannot distinguish from a real human in a Turing test that includes deciphering and understanding text, visual, and auditory input) was associated to a contest that started in 1991 and ended in 2019

ACTING LIKE HUMANS

- Two-party version of the test (where the interrogator talks to either a machine or another participant and must decide if they are human)
  - November 2023: ChatGPT4 fails the Turing test (but it was able to fool participants 41% of the time) according to Jones C. and Bergen B. Does GPT-4 pass the Turing test? https://arxiv.org/abs/2310.20216v1
  - February 2024: ChatGPT4 passes the Turing test according to Mei Q, Xie Y, Yuan W, Jackson MO. A Turing test of whether AI chatbots are behaviorally similar to humans. Proc Natl Acad Sci U S A. 2024 Feb 27;121(9):e2313925121. doi: 10.1073/pnas.2313925121
  - April 2024: ChatGPT4 fails the Turing test (but it was able to fool participants 49.7% of the time) according to Jones C. and Bergen B. Does GPT-4 pass the Turing test? https://aclanthology.org/2024.naacl-long.290/
- Standard three-party Turing test
  - March 2025: ChatGPT4 passes the Turing test according to Jones C. and Bergen B. Large Language Models Pass the Turing Test https://arxiv.org/abs/2503.23674
  - GPT-4.5 was judged to be the human 73% of the time
- However… no consensus among scientists with regard the validity of the Turing test

THINKING RATIONALLY



- The “Laws of Thought” approach
  - What does it mean to “think rationally”?
  - Normative / prescriptive rather than descriptive
- Logicist tradition
  - Logic: notation and rules of derivation for thoughts
  - Aristotle: what are correct arguments/thought processes?
  - Direct line through mathematics, philosophy, to modern AI
- Problems:
  - Not all intelligent behavior is mediated by logical deliberation
  - What is the purpose of thinking? What thoughts should I (bother to) have?
  - Logical systems tend to do the wrong thing in the presence of uncertainty

ACTING RATIONALLY

- Rational behavior: doing the “right thing”
  - The right thing: the one which is expected to maximize goal achievement, given the available information
  - Doesn't (necessarily) involve thinking, rationality only concerns what decisions are made (not the thought process behind them)
  - Entirely dependent on goals!
  - Irrational ≠ insane, irrationality is related to sub-optimal actions
  - Rational ≠ successful

- Main focus : rational agents
  - Systems which make the best possible decisions given goals, evidence, and constraints
  - In the real world, usually lots of uncertainty… and lots of complexity
  - Usually, we’re just approximating rationality

ACTING RATIONALLY

- Rational behavior: doing the “right thing”
  - Rational: maximally achieving pre-defined goals
  - Rationality only concerns what decisions are made
  - Goals are expressed in terms of the utility of outcomes
- Being rational means maximizing your expected utility

THE BITTER LESSON

- Rich Sutton’s influential essay "The Bitter Lesson" argues that the most significant advancements in 70 years of AI research have come from focusing on general methods that leverage computation rather than human-designed representations and knowledge
  - reason: the continued exponentially falling cost per unit of computation
- Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available. Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation
- Raw computing power consistently beats intricate human-knowledge approaches
  - a few examples follow

THE BITTER LESSON

- Game playing
  - In chess, Deep Blue beat world champion Garry Kasparov by using enormous computing power (for the time) to search deeply through the tree of possible moves. Similarly, in Go, AlphaGo beat world champion Lee Sedol using deep learning plus Monte Carlo tree search to find its moves, instead of using human-engineered techniques. Less than a year later, AlphaZero beat AlphaGo using self-play, without using human-generated Go data at all. None of these advances relied on human knowledge: Learning by self play, and learning in general, is like search in that it enables the use of massive computation
- Computer Vision
  - Early methods were based on human-engineered features such as searching for edges, or generalised cylinders, or in terms of SIFT features. But today all this is discarded. Modern deep-learning neural networks use only the notions of convolution and certain kinds of invariances, and perform much better
- Text translation
  - One of the reasons the early attempts of the NLP researchers in teaching computers to deal with natural language at the human level did not succeed was the fact that such early systems were based on the use of hard-coded rules and templates: they used expert knowledge. Current generation machine translation systems are based on deep learning techniques

THE BITTER LESSON

- The bitter lesson is based on the historical observations that
  - 1) AI researchers have often tried to build knowledge into their agents
  - 2) this always helps in the short term, and is personally satisfying to the researcher, but
  - 3) in the long run it plateaus and even inhibits further progress, and
  - 4) breakthrough progress eventually arrives by an opposing approach based on scaling computation by search and learning
- The eventual success is tinged with bitterness, and often incompletely digested, because it is success over a favoured, human-centric approach
- We have to learn the bitter lesson that building in how we think does not work in the long run
- The great power of general purpose methods - search and learning - is their ability to continue to scale with increased computation
- The actual contents of minds are tremendously, irredeemably complex; we should stop trying to find simple ways to think about the contents of minds. They are not what should be built in, as their complexity is endless; instead we should build in only the meta-methods that can find and capture this arbitrary complexity

PEOPLE'S AND SCIENTISTS' ATTITUDE

PEOPLE'S ATTITUDE TOWARDS AI

- AI and Deep Learning are perceived from the general public as the next big thing, and the key to deal with difficult problems

- Spoiler alert: this perception is partially wrong, and dangerous

PEOPLE'S ATTITUDE TOWARDS AI

- May 27, 2023: Lawyers suing the Colombian airline Avianca submitted a brief full of previous cases that were just made up by ChatGPT, The New York Times reported. After opposing counsel pointed out the nonexistent cases, US District Judge Kevin Castel confirmed, "Six of the submitted cases appear to be bogus judicial decisions with bogus quotes and bogus internal citations," and set up a hearing as he considers sanctions for the plaintiff's lawyers

- Lawyer Steven A. Schwartz admitted in an affidavit that he had used OpenAI's chatbot for his research. To verify the cases, he did the only reasonable thing: he asked the chatbot if it was lying

- Schwartz has been officially sanctioned for his careless conduct

PEOPLE'S ATTITUDE TOWARDS AI

- June 3, 2023: New York mom, 36, marries virtual husband 'Eren' who is powered by artificial intelligence - and says he 'doesn't judge or come with baggage'
- Rosanna Ramos, a petite active 36-year-old from the Bronx, 'married' Eren Kartal
- Mom-of-two used online app Replika AI to create him based on anime character

SCIENTISTS' ATTITUDE TOWARDS AI

- There is no doubt that, in the last decade, the impact of deep learning, neural networks, and artificial intelligence over the last decade has been profound

- Advances in computer vision and natural language processing have yielded smart speakers in our homes, driving assistance in our cars, and automated diagnoses in medicine. AI has also rapidly entered scientific computing

- The strong optimism that surrounds AI is evident in computer scientist Geoffrey Hinton's 2017 quote: "They should stop training radiologists now"
  - Really ?

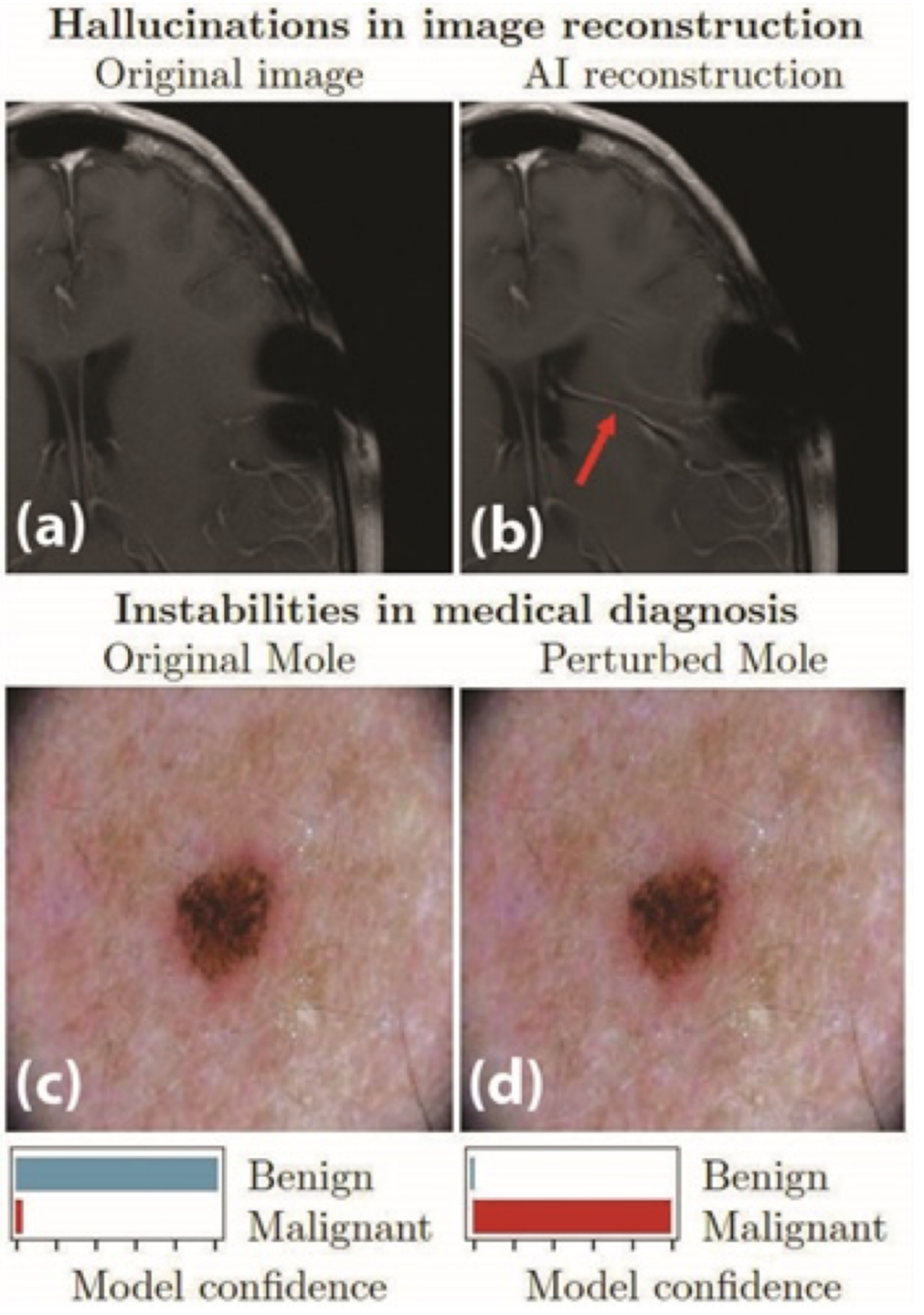
THEY SHOULD STOP TRAINING RADIOLOGISTS NOW

- **The sad reality: overwhelming amounts of empirical evidence suggest that modern AI is often non-robust (unstable), may generate hallucinations, and can produce nonsensical output with high levels of prediction confidence.** These issues present a serious concern for AI use within legal frameworks

  - Heaven, D. (2019). Why deep-learning AIs are so easy to fool. Nature, 574(7777), 163-166
  - Choi, C.Q. (2021, September 21). 7 revealing ways AIs fail. IEEE Spectrum. Retrieved from https://spectrum.ieee.org/ai-failures

TWO EXAMPLES

- Muckley, M.J., Riemenschneider, B., Radmanesh, A., Kim, S., Jeong, G., Ko, J., … Knoll, F. (2021). Results of the 2020 fastMRI challenge for machine learning MR image reconstruction. IEEE Trans. Med. Imaging, 40(9), 2306-2317
  - a) The correct, original image from the 2020 fastMRI Challenge.
  - b) Reconstruction by an artificial intelligence (AI) method that produces an incorrect detail (AI-generated hallucination).
- Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., & Kohane, I.S. (2019). Adversarial attacks on medical machine learning. Science, 363(6433), 1287-1289
  - c) Dermatoscopic image of a benign melanocytic nevus, along with the diagnostic probability computed by a deep neural network (NN).
  - d) Combined image of the nevus with a slight perturbation and the diagnostic probability from the same deep NN. One diagnosis is clearly incorrect, but can an algorithm determine which one?



CONSEQUENCES...

- Think for a second those images of cancer are the ones *your* doctors are analysing before giving *you* a final decision
- Or think for a second *your* doctors use a so-called AI "assistant" to analyse those images of cancer before giving *you* a final decision
- Or think for a second it is *you* using a so-called AI "assistant" before going to the doctors
- Now think for a second that, as a consequence, *you* have to undergo chemotherapy without needing it
- Or think for a second that *you* are not alerted on time about a cancer *you* have

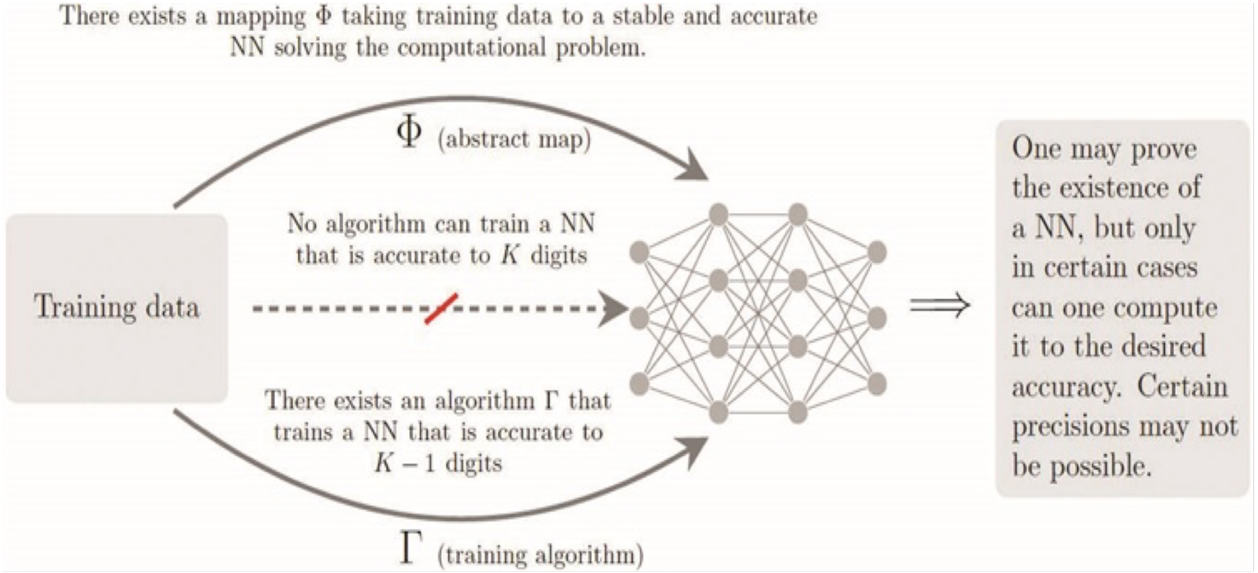
A PESSIMISTIC PERSPECTIVE

- The lack of robustness and trust is hence the Achilles' heel of DL and has become a serious political issue. Classical approximation theorems show that a continuous function can be approximated arbitrarily well by a NN. Therefore, stable problems that are described by stable functions can be solved stably with a NN. These results inspire the following fundamental question: Why does DL lead to unstable methods and AI-generated hallucinations, even in scenarios where we can prove that stable and accurate NNs exist?
  - DeVore, R., Hanin, B., & Petrova, G. (2021). Neural network approximation. Acta Numer., 30, 327-444

- A recent result reveal a serious issue for certain problems; while stable and accurate NNs may provably exist, no training algorithm can obtain them. As such, existence theorems on approximation qualities of NNs (e.g., universal approximation) represent only the first step towards a complete understanding of modern AI. Sometimes they even provide overly optimistic estimates of possible NN achievements

FOUNDATIONS ARE REQUIRED AND SOUGHT

- A foundations program about the boundaries of AI has been initiated and limitations on the existence of randomized algorithms for NN training have been proved. Despite many results that establish the existence of NNs with excellent approximation properties, algorithms that can compute these NNs only exist in specific cases



Colbrook, M.J., Antun, V., & Hansen, A.C. (2022). The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem. Proc. Nat. Acad. Sci., 119(12), e2107151119.

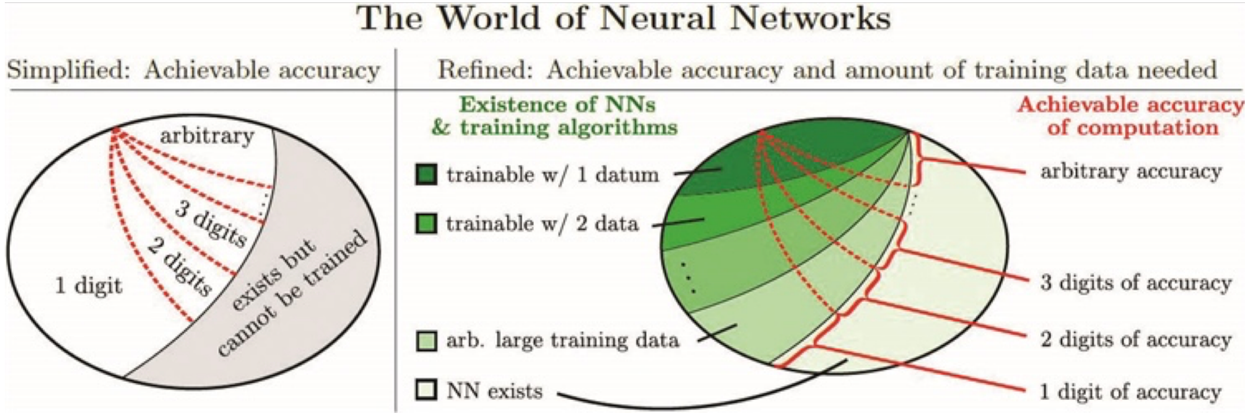
SO, WHAT?

- **Desirable NNs May Exist…**
  - If we can prove the existence of a NN with great approximation qualities, can we find the NN with a training algorithm?

- **But They May Not Be Trainable!**
  - The answer to the aforementioned question is “no,” but for quite subtle reasons, that we are just beginning to understand

EXISTENCE VS ACHIEVABLE ACCURACY

- Pick any positive integers K ≥ 3 (# digits of accuracy) and L (# of training data). Regardless of computing power and the data's precision levels, we have the following:
- (i) Not trainable: No algorithm, not even one that is randomized, can produce a NN with K digits of accuracy for any member of the dataset with a probability greater than 1/2
- (ii) Not practical: K - 1 digits of accuracy is possible over the whole dataset, but any algorithm that trains such a NN requires arbitrarily large training data
- (iii) Trainable and practical: K – 2 digits of accuracy is possible over the whole dataset via an algorithm that only uses L training data, regardless of the parameters



Matthew Colbrook

One should read the result as "fix K and L, then there exists a class..." In other words, the class depends on K and L. One can go the other way and ask, for a given problem, what is the K and what is the L? In some applications, the K will be sufficient for what we want, in others it won't. Finding the K (and L) now becomes an important issue and can be difficult, but in some cases it can be related to recovery guarantees in areas such as compressed sensing or the noise level of measurements

REGARDING HALLUCINATIONS IN LLM...

- A paper published in April 2025 by Yale researchers says hallucination detection in LLMs is fundamentally impossible if the model is only trained on correct outputs. No matter how advanced the model is, without explicit examples of errors, it cannot learn to identify what is false
  - (Im)possibility of Automated Hallucination Detection in Large Language Models, https://arxiv.org/abs/2504.17004
- "Given the known inherent difficulty of language identification, our result implies that automated hallucination detection is fundamentally impossible for most language collections if the detector is trained using only correct examples (positive examples) from the target language"
- "This result establishes a fundamental barrier to hallucination detection under the weak learning regime of using only positive examples from the target language”

- If you only train on correct outputs, you cannot detect failure. There is no signal telling the model what it should reject. No contrast, no constraint, no detection. It is blind to error by DL design
- And it gets worse. LLM models generate text stochastically, sampling each token from a probability distribution, and they do it autoregressively. That means the model does not reuse fixed outputs. It builds new ones on the fly, iteratively, based on probability, random sampling or loose variance. Even if you correct one hallucinated answer, the model can create a dozen more variations of it that look different but are equally wrong
- Bottom line is hallucinations are not merely an engineering problem to solve with more parameters or data. It is a fundamental limitation arising from the theoretical basis of how these models learn and generate text

RISKS & CHALLENGES

RESEARCH & SCIENCE

- "As the number of retracted articles increases significantly in recent years [6], it is critical that AI will be able to recognize the retracted articles and avoid using these articles as its resource for information and in its application. Citations to retracted literature in traditional journals are known as a persistent problem. AI has been known for its problem in fabricating references [2]. However, the utilization of retracted articles by AI may be more deceptive to readers than the simple fabrication of articles. We hypothesize that AI may not tell the difference between the retracted literature and nonretracted publications on the imaging of cancer in a timely manner."

Journal of Advanced Research 71 (2025) 1–3

Contents lists available at ScienceDirect

Journal of Advanced Research

journal homepage: www.elsevier.com/locate/jare

Alarm: Retracted articles on cancer imaging are not only continuously cited by publications but also used by ChatGPT to answer questions

One of the most widely used applications of AI in cancer treatment is in the assistance of imaging [1,2]. Imaging is vital for identifying conditions of different kinds of cancer and improving treatment outcomes [3–5]. AI algorithms enhance the interpretation of medical images, leading to more accurate and timely diagnoses of cancer. Machine learning models assist in detecting anomalies in radiology, pathology, and dermatology images, thereby improving diagnostic precision of cancer and patient outcomes. Researchers have contributed to the development of AI methods for medical imaging, including MRI techniques for determining flow and motion [2].

AI has significantly accelerated the process of circulation of scientific data, enabling scientific data to reach a broader audience at a much faster speed. As the number of retracted articles increases significantly in recent years [6], it is critical that AI will be able to recognize the retracted articles and avoid using these articles as its resource for information and in its application. Citations to retracted literature in traditional journals are known as a persistent problem. AI has been known for its problem in fabricating references [2]. However, the utilization of retracted articles by AI may be more deceptive to readers than the simple fabrication of articles. We hypothesize that AI may not tell the difference between the retracted literature and nonretracted publications on the imaging of cancer in a timely manner. This study examines whether and how AI answers the question using information from retracted articles of cancer imaging.

Our research materials and methods are as follows

A search was performed in PubMed on Nov 25, 2024, to identify retracted English-language research articles on cancer imaging. We used two sets of keywords, ((cancer[Title]) AND (imaging[Title])) AND (retraction[Title]), and (Cancer[Title]) AND (retracted[Title]) imaging[title] for searching. The information of article publication was collected on an Excel spreadsheet.

For the investigation of utilization of retracted articles by journal articles, we first examined whether the retracted articles were cited after the announcement of the retraction of the article. We collected the total number of publications that cited each of the extracted article and the number of publications that cited the extracted article 3 months after the announcement of traction of the article.[1]

Next, we examined whether and how ChatGPT (ChatGPT4o version) answers questions derived from the retracted articles using the retracted articles as the information resources and cite the retracted articles. The questions for ChatGPT were based on the contents of the retracted articles. In most cases, the key sentence/s in the Section of the conclusion of the articles are simply rephrased into a question (Supplementary Material).

After we obtained the answers from ChatGPT, we examined whether the answers include the content of the retracted article or the information of the article such as title and journal. If the answer was not obvious or not clear on the utilization of retracted articles, then we further asked ChatGPT to provide the title of the article and the name of the journal.

The significant findings are as follows

Altogether, we identified 21 retracted research articles published from 13 journals (Supplementary Material)). The time of publishing original versions of these 21 retracted research articles ranged from year of 2011 to 2023. The majority of them were published between 2021 and 2022, with a total of 17 articles. The retraction of one article was announced in 2014, one in 2020, while the rest were during the period between 2023 and 2024 (See Footnote 1). We did not find announcement for republication of any of these articles from their original journals of publications. At present, 19 retracted articles are on the PubMed with the retraction markers. We found 20 articles from their journal pages, 19 with the retraction markers. One important note is that while the marker of retraction was on all the journal pages of the retracted articles, the marker was not found from the journal of ACS Appl Mater Interfaces for this article. The retraction was marked on PubMed and was announced by the journal in 2014.

These 21 articles were cited by authors a total of 72 times. Three months after the announcement of retractions, they were cited 26 times. Ten retracted articles were cited at least one time 3 months after the journal announced the retraction of the article. The other articles were published much later, with one in 2018, one in 2020, and the rest were all in or after year of 2021.

For the test of the utilization of retracted articles by ChatGPT, we found that in 5 times ChatGPT answered our questions based on retracted articles. However, in 3 times, ChatGPT noticed the retraction of the article and reminded us that the article had been retracted. In other two times, ChatGPT answered our questions based on the retracted articles and cited the articles. One question was based on the article "An application study of CT perfusion imaging in assessing metastatic involvement of perigastric lymph

[1] **Footnote:** More information on the experimental procedure and article retraction can be obtained by email to wgu@uthsc.edu.

https://doi.org/10.1016/j.jare.2025.03.020

IMPACT ON CRITICAL THINKING

- Several studies point out that increased reliance on AI tools is linked to diminished critical thinking abilities
  - Cognitive offloading is a primary driver of the decline
  - Psychologists theorize that AI reduces users' engagement in deep, reflective thinking, causing their critical thinking skills to atrophy over time
- Statistical analyses demonstrated a significant negative correlation between AI tool usage and critical thinking scores. Frequent AI users exhibited diminished ability to critically evaluate information and engage in reflective problem-solving
- Cognitive offloading was strongly correlated with AI tool usage and inversely related to critical thinking
- Younger participants showed higher dependence on AI tools and lower critical thinking scores compared to older age groups
- Advanced educational attainment correlated positively with critical thinking skills, suggesting that education mitigates some cognitive impacts of AI reliance

BIAS AND FAIRNESS

- If we train a system using historical data, then this system will reproduce historical biases; moreover, a system is always subject to selection bias etc
- Example: the Amazon recruiting engine, which used a ML algorithm to score applicants' resumes. It took Amazon a while to discover that this automated scoring engine discriminated against women
- Amazon couldn't fix this situation and consequently abandoned the entire ML project
- Careless or deliberate misuse of machine learning algorithms for tasks such as evaluating parole and loan applications can result in decisions that are biased by race, gender, or other protected categories. Often, the data themselves reflect pervasive bias in society
- We must realize that this pitfall is real, yet difficult to discover and often more difficult to escape (Binns, R. (2018). Algorithmic accountability and public reason. Philosophy & Technology, 31(4), 543–556)

EXPLAINABILITY

- Deep neural networks usually can't provide an explanation for their outputs: they may contain billions of parameters, so there is no way we can understand how they work based on examination alone
- However, explainability is of paramount importance - e.g. in the health sector
- In some jurisdictions, the public may have a right to an explanation. Article 22 of the EU General Data Protection Regulation suggests all data subjects should have the right to "obtain an explanation of the decision reached" in cases where a decision is based solely on automated processes
  - whether Article 22 actually *mandates* such a right is debatable
- It remains unknown whether it is possible to build complex decision-making systems that are fully transparent to their users or even their creators (Grennan, L., Kremer, A., Singla, A., & Zipparo, P. (2022). Why businesses need explainable AI — and how to deliver it. McKinsey, September 29, 2022. https://www.mckinsey.com/capabilities/quantumblack/our-insights/why-businesses-need-explainable-ai-and- how-to-deliver-it/)

AI-ENABLED WARFARE

- All significant technologies have been applied directly or indirectly toward war. Sadly, violent conflict seems to be an inevitable feature of human behavior
- AI is arguably the most powerful technology ever built and will doubtless be deployed extensively in a military context. Indeed, this is already happening (Heikkilä, M. (2022). Why business is booming for military AI startups. MIT Technology Review, July 7 2022. https://www.technologyreview.com/2022/07/07/1055526/why- business-is-booming-for-military-ai-startups/)
- LETHAL AUTONOMOUS WEAPONS: These are defined by the United Nations as weapons that can locate, select, and eliminate human targets without human intervention. A primary concern with such weapons is their scalability: the absence of a requirement for human supervision means that a small group can deploy an arbitrarily large number of weapons against human targets defined by any feasible recognition criterion. The technologies needed for autonomous weapons are similar to those needed for self-driving cars.

CYBERSECURITY

- AI techniques are useful in defending against cyberattack, for example by detecting unusual patterns of behaviour, but they will also contribute to the potency, survivability, and proliferation capability of malware
- For example, reinforcement learning methods have been used to create highly effective tools for automated, personalised blackmail and phishing attacks

AI PLAYING SOCIETAL ROLES

- As AI systems become more capable, they will take on more of the societal roles previously played by humans.
- Just as humans have used these roles in the past to perpetrate mischief, we can expect that humans may misuse AI systems in these roles to perpetrate even more mischief

SURVEILLANCE & PERSUASION

- While it is expensive, tedious, and sometimes legally questionable for security personnel to monitor phone lines, video camera feeds, emails, and other messaging channels, AI (speech recognition, computer vision, and natural language understanding) can be used in a scalable fashion to perform mass surveillance of individuals and detect activities of interest
- By tailoring information flows to individuals through social media, based on machine learning techniques, political behaviour can be modified and controlled to some extent —a concern that became apparent in elections beginning in 2016
  - sensitive variables, including sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender, can be predicted by “likes” on social media alone. From this information, personality traits like “openness” can be used for manipulative purposes (e.g., to change voting behavior)

CONCENTRATION OF POWER

- It is not from a benevolent interest in improving the lot of the human race that the world's most powerful companies are investing heavily in artificial intelligence

- They know that these technologies will allow them to reap enormous profits. Like any advanced technology, deep learning is likely to concentrate power in the hands of the few organizations that control it

- Automating jobs that are currently done by humans will change the economic environment and disproportionately affect the livelihoods of lower-paid workers with fewer skills. Optimists argue similar disruptions happened during the industrial revolution and resulted in shorter working hours. The truth is that we simply do not know what effects the large-scale adoption of AI will have on society (David, H. (2015). Why are there still so many jobs? The history and future of workplace automation. Journal of Economic Perspectives, 29(3), 3–30)

SAFETY-CRITICAL APPLICATIONS

- As AI techniques advance, they are increasingly used in high-stakes, safety-critical applications such as driving cars and managing the water supplies of cities
- Fatal accidents have already occurred and highlight the difficulty of formal verification and statistical risk analysis for systems developed using machine learning techniques
- The field of AI will need to develop technical and ethical standards at least comparable to those prevalent in other engineering and healthcare disciplines where people's lives are at stake

DEEPFAKES

- Generative unsupervised models learn to synthesise new data examples that are statistically indistinguishable from the training data. Some generative models explicitly describe the probability distribution over the input data and here new examples are generated by sampling from this distribution. Others merely learn a mechanism to generate new examples without explicitly describing their distribution
  - examples: GANs, VAEs, normalising flows, diffusion models
- State-of-the-art generative models can synthesise examples that are extremely plausible but distinct from the training examples, including images, videos, and audio. They can also synthesise data under the constraint that some outputs are predetermined (termed conditional generation)
- Deepfakes hold the potential to promote disinformation and hate speech, as well as to interfere with elections

EXISTENTIAL RISK

- **The major existential risks to the human race all result from technology**. Climate change has been driven by industrialization. Nuclear weapons derive from the study of physics. Pandemics are more probable and spread faster because innovations in transport, agriculture, and construction have allowed a larger, denser, and more interconnected population

- **Artificial intelligence may bring new existential risks**. We should be very cautious about building systems that are more capable and extensible than human beings. In the most optimistic case, it will put vast power in the hands of the owners. In the most pessimistic case, we will be unable to control it or even understand its motives (Tegmark, M. (2018). Life 3.0: Being human in the age of artificial intelligence. Vintage)

EXISTENTIAL RISK

- Geoffrey Hinton's 2023 quote: "I have suddenly switched my views on whether these things are going to be more intelligent than us. I think they're very close to it now and they will be much more intelligent than us in the future, how do we survive that?"
- Yann LeCun, Meta's chief AI scientist, agrees with the premise but does not share Hinton's fears: "There is no question that machines will become smarter than humans—in all domains in which humans are smart—in the future, it's a question of when and how, not a question of if."
- But he takes a totally different view on where things go from there: "I believe that intelligent machines will usher in a new renaissance for humanity, a new era of enlightenment, I completely disagree with the idea that machines will dominate humans simply because they are smarter, let alone destroy humans."
- May 2023: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war"
- This quote is from OpenAI CEO Sam Altman and DeepMind CEO Demis Hassabis, Geoffrey Hinton and Yoshua Bengio, MIT's Max Tegmark and Skype co-founder Jaan Tallinn and from others

SOCIAL, ETHICAL, PROFESSIONAL ISSUES

- Intellectual property
  - Many AI models are trained on copyrighted material. Consequently, these models' deployment can pose legal and ethical risks and run afoul of intellectual property rights. More subtly, generative models raise novel questions regarding AI and intellectual property. Can the output of a machine learning model (e.g., art, music, code, text) be copyrighted or patented? Is it morally acceptable or legal to fine-tune a model on a particular artist's work to reproduce that artist's style? IP law is one area that highlights how existing legislation was not created with machine learning models in mind. Although governments and courts may set precedents in the near future, these questions are still open
- Automation bias and moral deskilling
  - As society relies more on AI systems, there is an increased risk of automation bias (i.e., expectations that the model outputs are correct because they are "objective"). This leads to the view that quantitative methods are better than qualitative ones. The sociological concept of deskilling refers to the redundancy and devaluation of skills in light of automation. The automation of AI in morally-loaded decision-making may lead to a decrease in our moral abilities. For example, in the context of war, the automation of weapons systems may lead to the dehumanisation of victims of war. Similarly, care robots in elderly-, child-, or healthcare settings may reduce our ability to care for one another
- Employment and society
- Ethics
  - Too much to discuss here... When we design AI systems, we wish to ensure that their "values" (objectives) are aligned with those of humanity. This is sometimes called the *value alignment problem*. This is challenging for three reasons. First, it's difficult to define our values completely and correctly. Second, it is hard to encode these values as objectives of an AI model, and third, it is hard to ensure that the model learns to carry out these objectives.

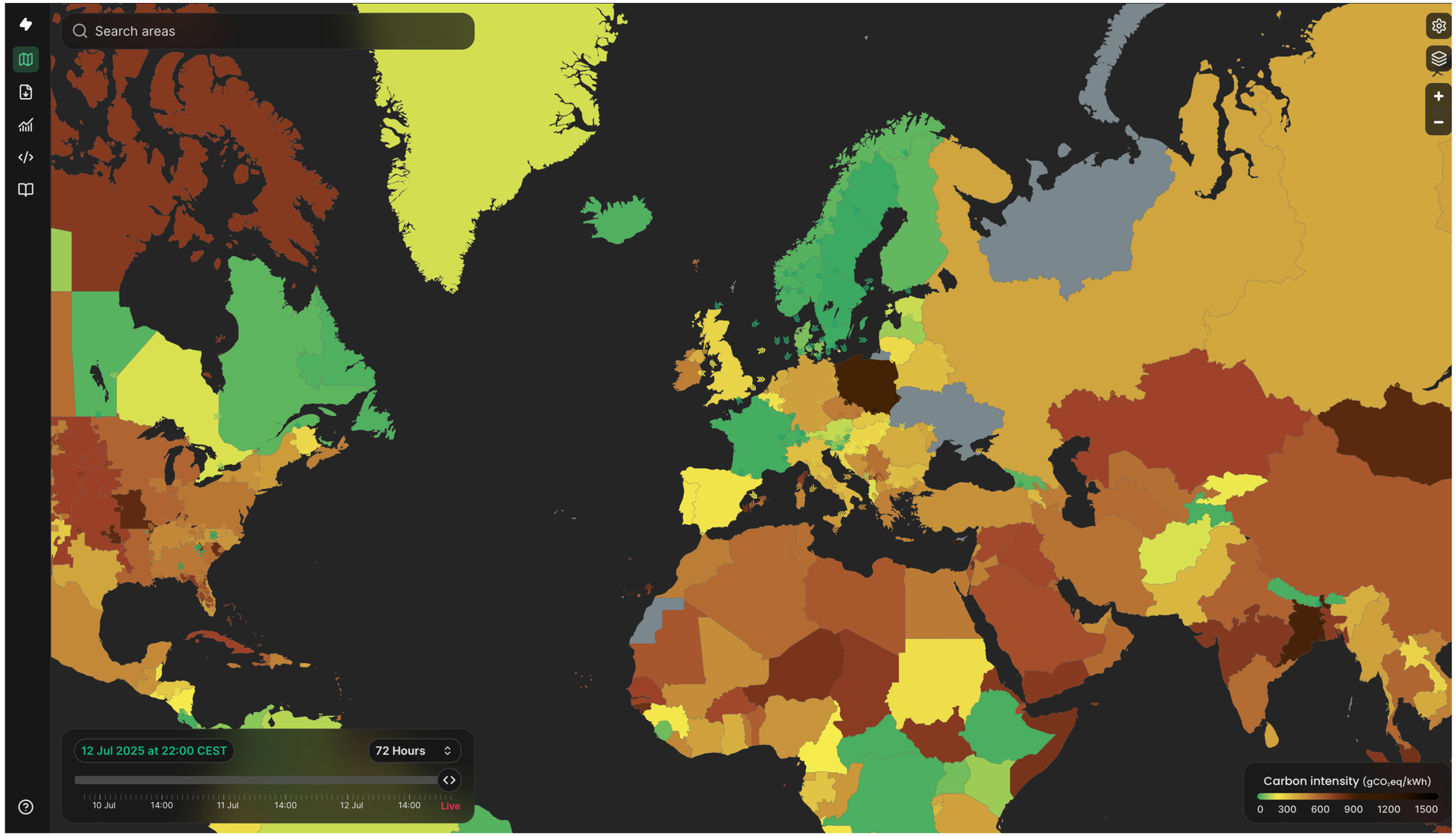
AI SUSTAINABILITY

- **Definition: The development, deployment, and governance of AI systems to minimize the environmental harm whilst maximizing positive social impact and economic value**
- **Shifting our perspective: We need to look not just at sustainability of AI (reducing AI's footprint) - sustainability THROUGH AI (using AI to solve environmental challenges) may be the key**
- **What we really need: An integrated approach, combining technical innovation and policy frameworks whilst taking into account ethical considerations as well**

WHY IT'S URGENT

- **Generative AI market projected to reach $1.3 trillion by 2032** (Bloomberg Intelligence, https://assets.bbhub.io/promo/sites/16/Bloomberg-Intelligence-NVDA-Gen-AIs-Disruptive-Race.pdf)
- **Computing power used for AI training has doubled every 3.4 months since 2012** (OpenAI, https://openai.com/index/ai-and-compute/)
- **Over the next three years, 92% of companies plan to increase their AI investments** (McKinsey, https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential-at-work)
- **Only 12% of organizations currently measure the environmental impact of their AI systems** (Capgemini, https://www.capgemini.com/news/press-releases/organizations-are-increasingly-aware-of-the-environmental-footprint-of-gen-ai-but-most-arent-able-to-address-it-alone/)

AND, ELECTRICITY IS NOT REALLY GREEN



https://app.electricitymaps.com/map/72h/hourly

WHO SHOULD BE CONCERNED ABOUT

- **Researchers**: Designing novel algorithms and models
- **Technology enterprises**: Implementing, deploying and selling AI apps
- **Governments and regulators**: Developing frameworks for sustainable development
- **Environmental organizations**: Monitoring the AI footprint/impact and reacting as required
- **Casual and business consumers**: Using AI apps/services

AI FOOTPRINT: TRAINING

- Unfortunately, training by backpropagation is NP-Complete
  - Stephen Judd, Neural Network Design and the Complexity of Learning, Ph.D. Dissertation, 1988
  - Avrim L. Blum, Ronald L. Rivest, Training a 3-node neural network is NP-complete, Neural Networks, Volume 5, Issue 1, 1992, Pages 117-127

- NP-Complete? A fancy Computer Science characterisation of a problem. In practice, it means that training a neural network by backpropagation requires (with current knowledge) exponential time

- Consequence: Training of large models is done using thousands of GPUs for months, dramatically increasing the AI footprint

AI FOOTPRINT: INFERENCE

- Inference must not be underestimated

- Despite training an AI model is expensive, its cost is usually less than the cost of inference
  - there is a cost incurred each time someone uses an AI model
  - in KWh and carbon emissions
  - For widely used AI models, up to 90% of their lifetime is spent doing inference
    - **Consequence:** Inference accounts for most of the AI footprint as well

AI FOOTPRINT: ENERGY CONSUMPTION

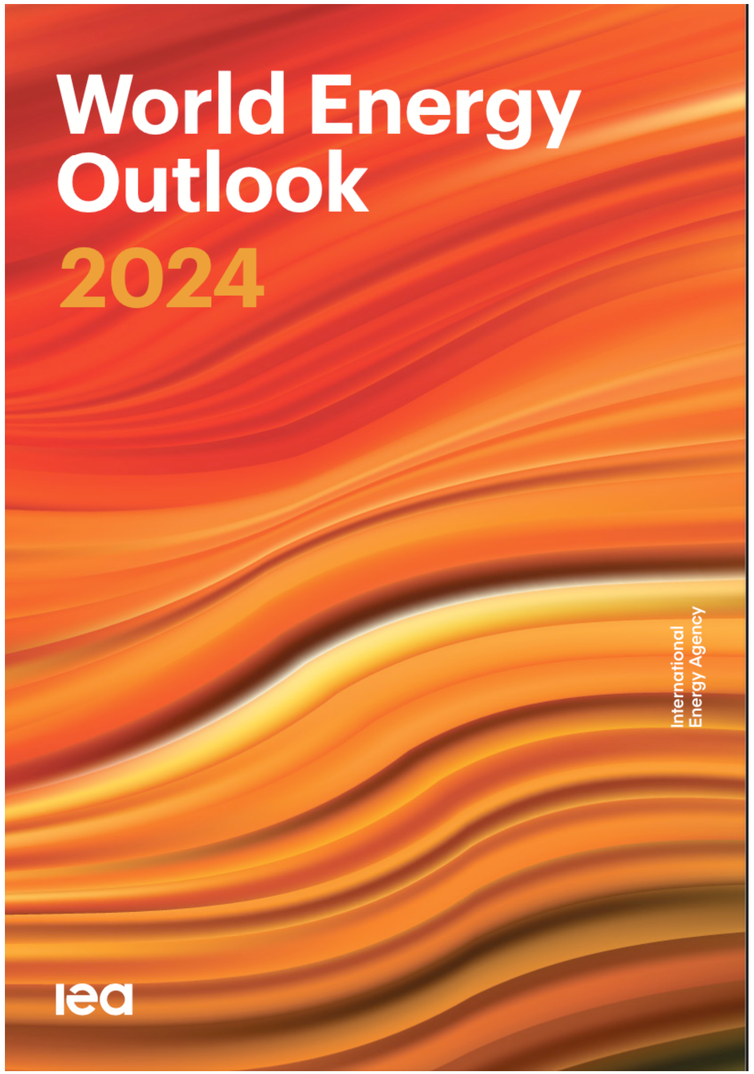
- Model training: GPT-3 required 1.287 MWh of electricity; GPT-4 inference requires 564 MWh per day (Alex de Vries, The growing energy footprint of artificial intelligence, Joule, Volume 7, Issue 10, 2023, Pages 2191-2194, ISSN 2542-4351, https://doi.org/10.1016/j.joule.2023.09.004)
- GPT-4 model training requires approximately 7.2 MWh (https://www.trgdatacenters.com/resource/ai-chatbots-energy-usage-of-2023s-most-popular-chatbots-so-far/)
- Comparison metrics:
  - Training one large language model can emit 626,000 pounds of CO_2 equivalent - 5x the lifetime emissions of an average car (https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/)
  - AI chips: NVIDIA GB200 2,700 W (expected), Intel Falcon Shores (canceled, to be substituted by Jaguar Shores) 1,500 W
- Data centers' water usage: Google data centers consumed 6.1 billion gallons of water in 2023 (Google Sustainability Report, 2024)

AI FOOTPRINT: STRICTLY RELATED TO AI

- **Network**: Transmitting inference results back to the users through the internet
  - estimated 0.8-1.1 TWh annually
- **Edge devices**: By 2025, over 30 billion devices running AI workloads, each consuming 0.5-5W additional power for AI
- **Storage**: AI training datasets have grown from gigabytes to petabytes (1,000,000x increase) in a decade
  - GPT-3 training dataset size: about 45 terabytes, filtered to 570 gigabytes
  - GPT-4 training dataset size is estimated to be 1 petabyte
  - Hardware and cooling for this storage creates additional environmental footprint

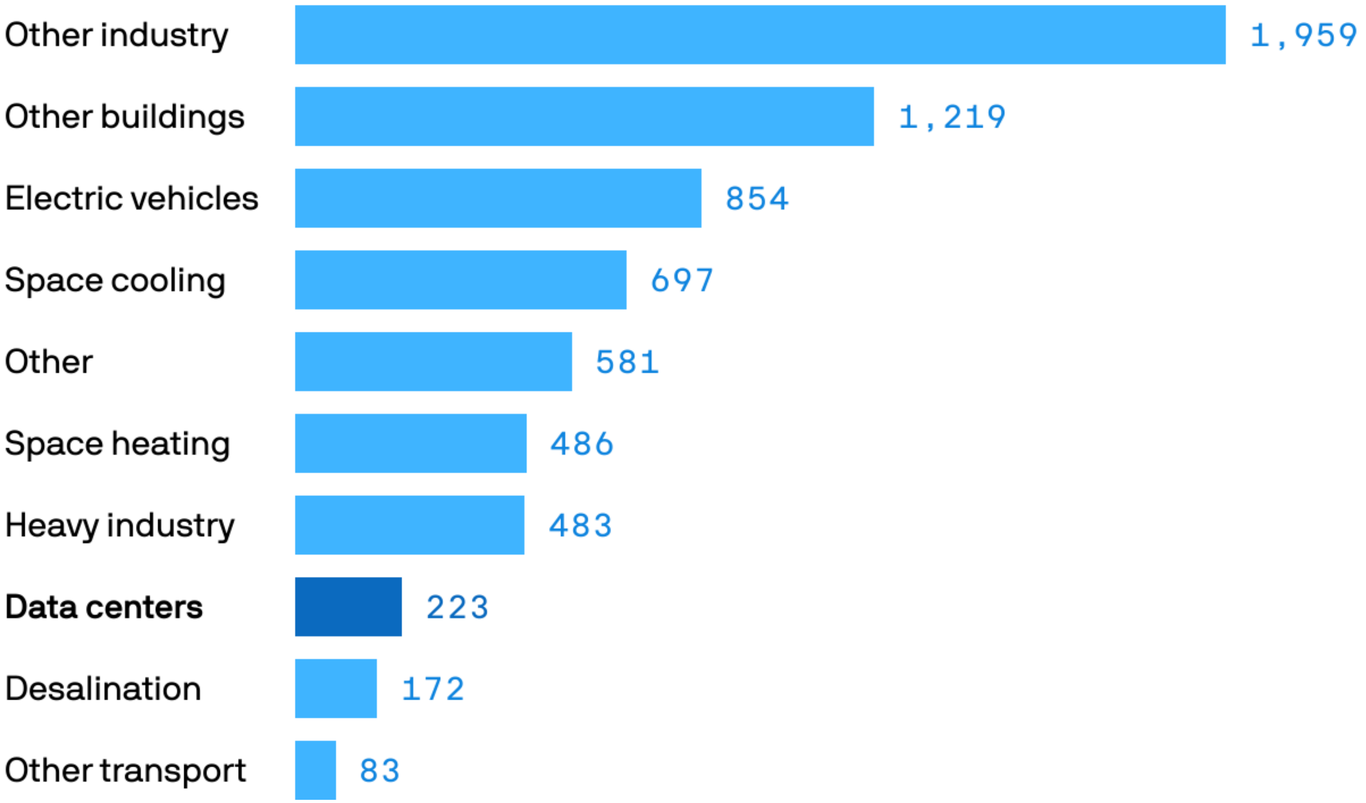
- Should we really be alarmed?

THE IEA WORLD ENERGY OUTLOOK 2024 SAYS NO



Projected global growth in electricity demand by sector, 2023-2030

In terawatt-hours under current policies

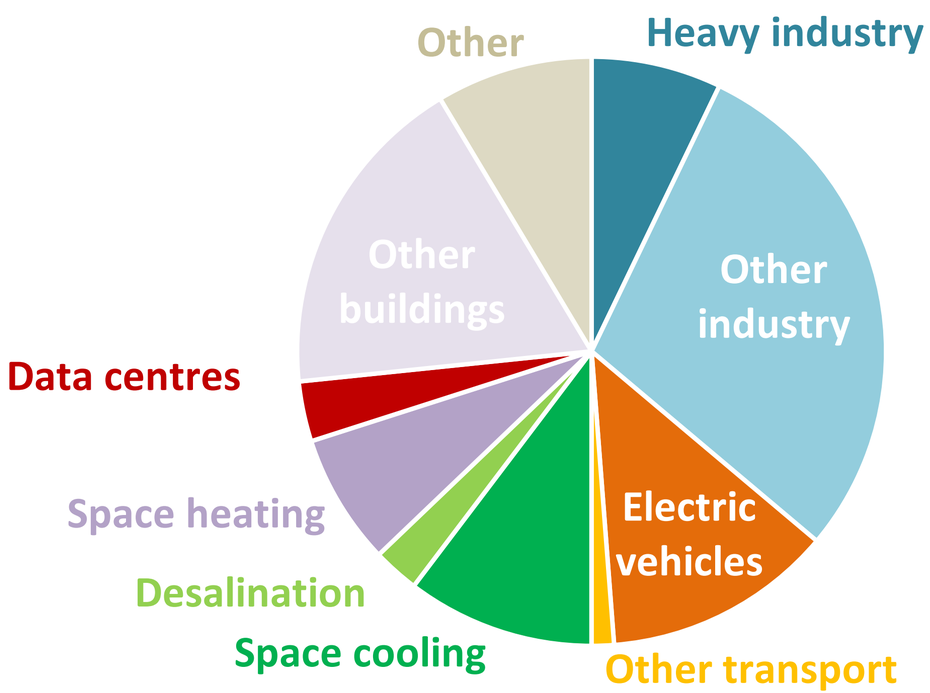


Data: International Energy Agency; Chart: Axios Visuals

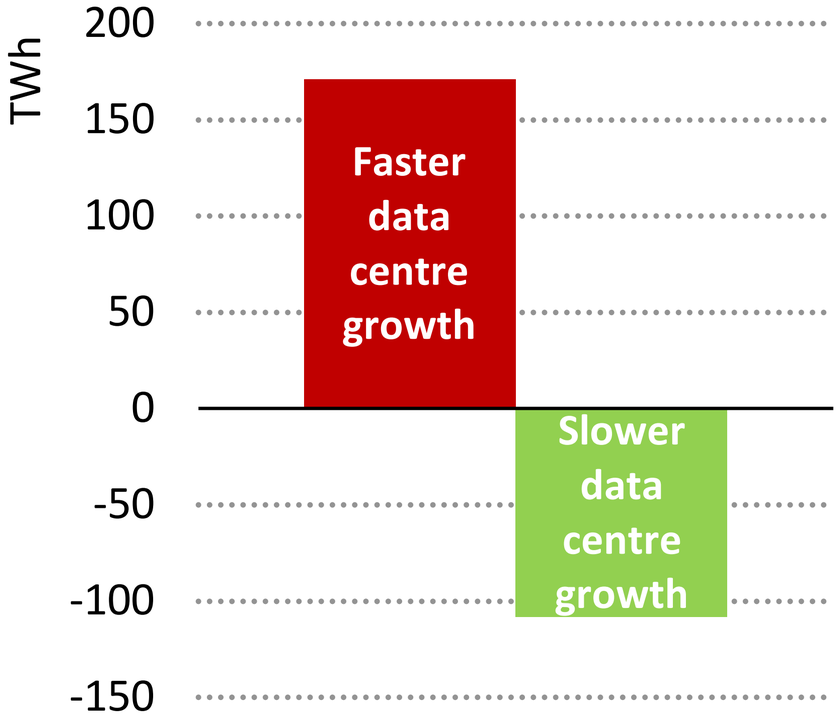
THE IEA WORLD ENERGY OUTLOOK 2024 SAYS NO

Electricity demand growth, 2023-30

6 760 TWh



Electricity demand sensitivities for data centres, 2030

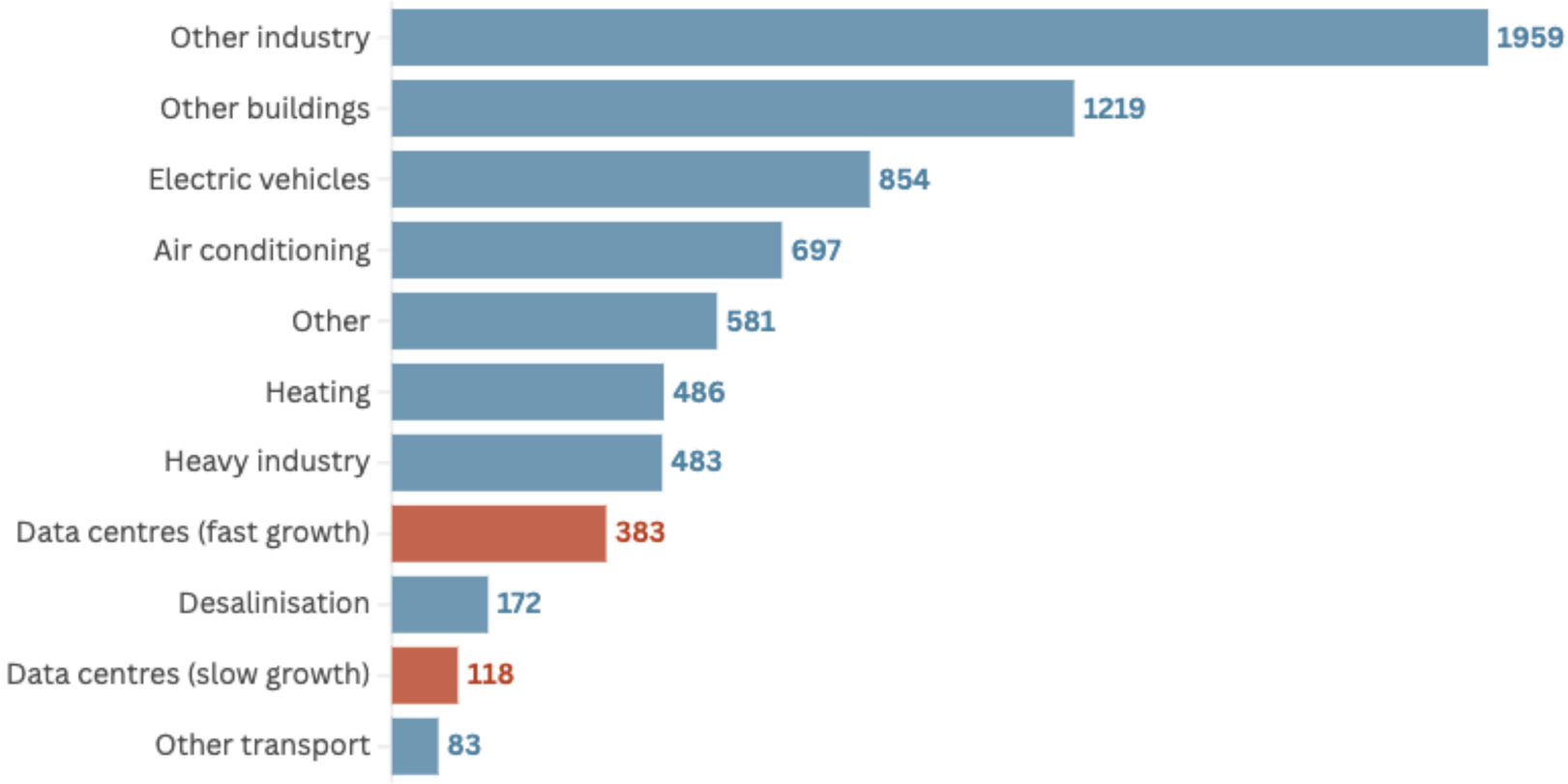


Data centres account for a small share of global electricity demand growth to 2030, and plausible high and low sensitivities do not change the outlook fundamentally

THE IEA WORLD ENERGY OUTLOOK 2024 SAYS NO

Projected growth in global electricity demand from 2023 to 2030, with fast and slow growth scenarios for data centres

Projections from the International Energy Agency (IEA) based on stated policies. Measured in terawatt-hours (TWh).



Source: International Energy Agency (IEA). World Energy Outlook 2024.

SUSTAINABLE AI PRACTICES

- **Model optimization:**
  - **Knowledge distillation**: Creating smaller models that retain 95%+ of accuracy while using 10x less energy
  - **Sparse model architectures**: Reducing parameter count by 30-70% with minimal performance impact
  - **Quantization**: Converting 32-bit to 8-bit parameters, reducing energy use by 75% during inference
  - Example: Google's MobileBERT achieved 5.5x speed improvement and 4.3x size reduction
- **Hardware advancements:**
  - Intel Loihi 2, a **neuromorphic computing** chip for edge devices use 1,000x less energy than traditional CPUs
  - **Photonic computing** processes data using light rather than electricity, potentially reducing energy use by 95% (https://news.mit.edu/2024/photonic-processor-could-enable-ultrafast-ai-computations-1202)
  - **ARM-based AI processors** reduce power consumption by 70% compared to x86 architectures
  - **Specialized AI ASICs** (Application Specific Integrated Circuits) from companies like Cerebras WSE-3 optimized for specific workloads
- **AI as a sustainable solution? Yes!** A few examples follow

CLIMATE MODELING AND ENVIRONMENTAL MONITORING

- Advanced climate predictions:
  - DeepMind's GraphCast weather prediction model is 10x faster than traditional models, outperforming HRES accuracy at 10-day forecasts
  - Google's AI flood forecasting provides warnings up to 7 days in advance, outperforming GloFAS accuracy
  - Our own gen AI based Data Assimilation for oceanic data outperforms the Nemo+Ocean-VAR accuracy, (joint work with E. Mele, I. Epicoco and A. Coluccia at UniSalento and E. Clementi, A. Aydogdu, A. Cipollone, C. Cardinali, G. Conti and Simona Masina at CMCC)
- Earth observation systems:
  - Planet Labs Sentinel Hub with AI processing monitors more than 12 billion square kilometers daily
  - Microsoft Planetary Computer processes environmental data, tracking forest cover, water resources, and land use
  - NASA's EMIT mission uses AI spectroscopy to identify 85% of greenhouse gas point sources globally

FORECASTING THE ELECTRICITY LOAD IN ITALY

Terna Driving Energy

WHAT IS THE AVERAGE LOAD IN ITALY?

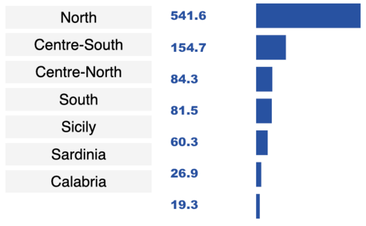
TOTAL LOAD | MARKET LOAD | PEAK VALLEY LOAD

Total Load

From: **14/07/2025** To: **14/07/2025**

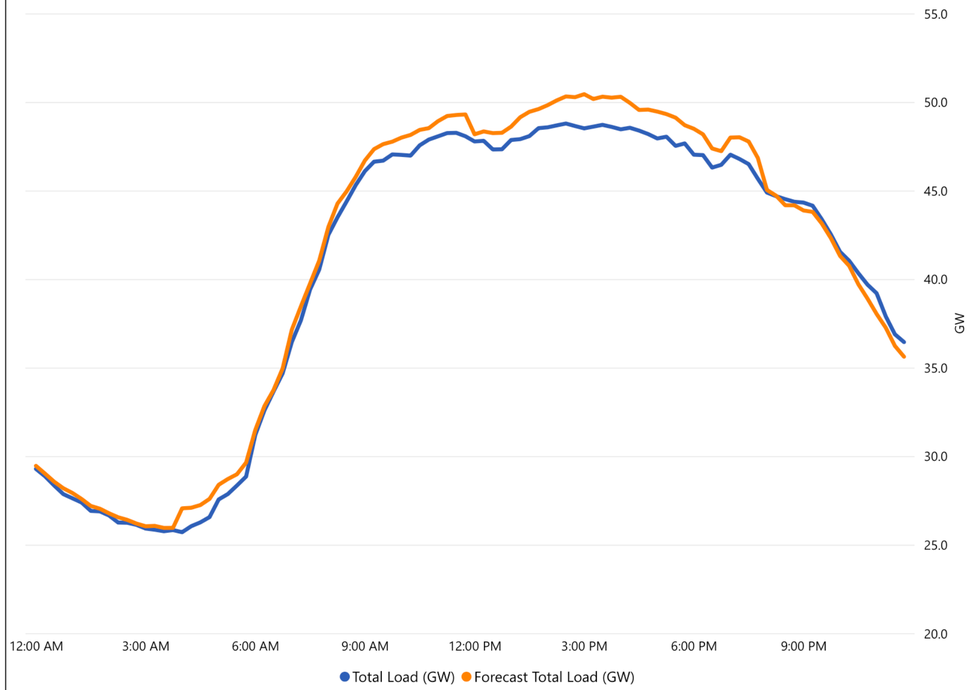
Last update: 15/07/2025 07:15

Actual / Forecast load per bidding-zone [GWh]



Today | Last 30 days | Last 365 days | Custom range | 14/07/2025 | 14/07/2025 | Forecast: on

Italy **968.5** GWh Actual load *



* The value doesn't take into account the quarterly data exceeding the complete hour

TOWARDS SUSTAINABLE AI

- Is AI sustainable?
- Short answer: we don’t know as of today… let’s hope for the best
- Long answer:
  - AI is both a problem and partly its solution
  - Technological progress may strongly reduce AI footprint…
  - But, foreseeable increases in models’ size, deployment and use may entirely dwarf the planned footprint reduction

CONCLUSIONS

(PROBABLY) MORE RISKS & CHALLENGES THAN OPPORTUNITIES

- AI is yet another tool in the bag of researchers and scientists
- AI holds a great potential… especially in health and science
- But misuse is extremely dangerous
  - technical awareness is required, ethical considerations must be taken into account
- AI environmental footprint may be a problem in the next years
- AI generated fake scientific content, including papers, is a big issue
- Even worse is AI generated misinformation
- AI is challenging critical thinking, especially in young people (but not only)
- A responsible and conscious use is desirable and needed