

CILC 2024
39th Italian Conference on Computational Logic

AI Risk and Neurosymbolic AI

Artur d'Avila Garcez
26 June 2024
City, University of London

1 billion LLM downloads (Hugging Face 2022-2024)

50%+ of downloads GPT2

ChatGPT made available to 100M users without safety review

LLMs reaching a plateau (GPT4 released more than 1 year ago)

Meta's US\$ 30 billion to buy 1M GPUs (May 2024)

- Most harmful things:

- Disinformation (elections), impersonation (identity theft), copyright violation (lawsuits), cheating at schools...

- Most beneficial things:

- Content summarization, language translation, content generation (creative industries), radiology (multimodal; AI applications to FDA up by 30%)...

ChatGPT: a solution without a problem?

- AI as *general-purpose* technology (Strong AI), productivity gains everywhere!
- General intelligence: adapting to novelty, planning, learning from few examples
- Narrow AI (domain specific): many relevant applications around the corner...
 - Healthcare: personalized cancer care, patient triage, drug discovery
 - Finance: robot advisors, fraud detection, risk prediction
 - Retail: logistics optimization, personal recommender system
 - Manufacturing: automation of production, distribution optimization
 - Transport: autonomous vehicles, route planning
 - Agriculture: precision agriculture, new materials for carbon capture
 - Education: virtual tutor, document summarization, language translation
 - Marketing: content creation, market segmentation
 - Legal services: virtual assistant, case law search automation
 - Energy: targeted demand response, accelerating fusion research

AI Index 2024 (HAI, Stanford University)

AI publications (% of total) by sector and geographic area, 2022

Source: Center for Security and Emerging Technology, 2023 | Chart: 2024 AI Index report

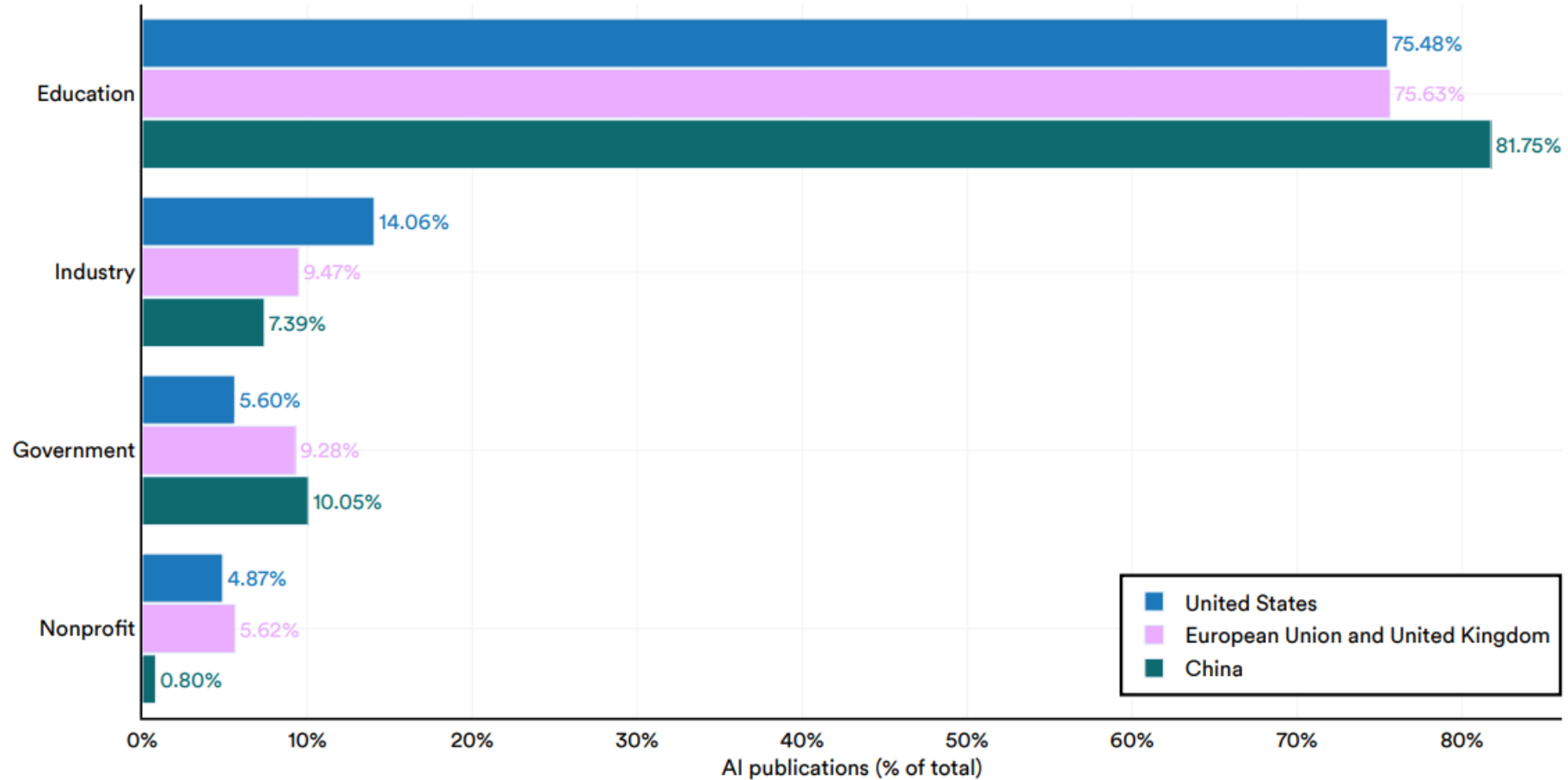
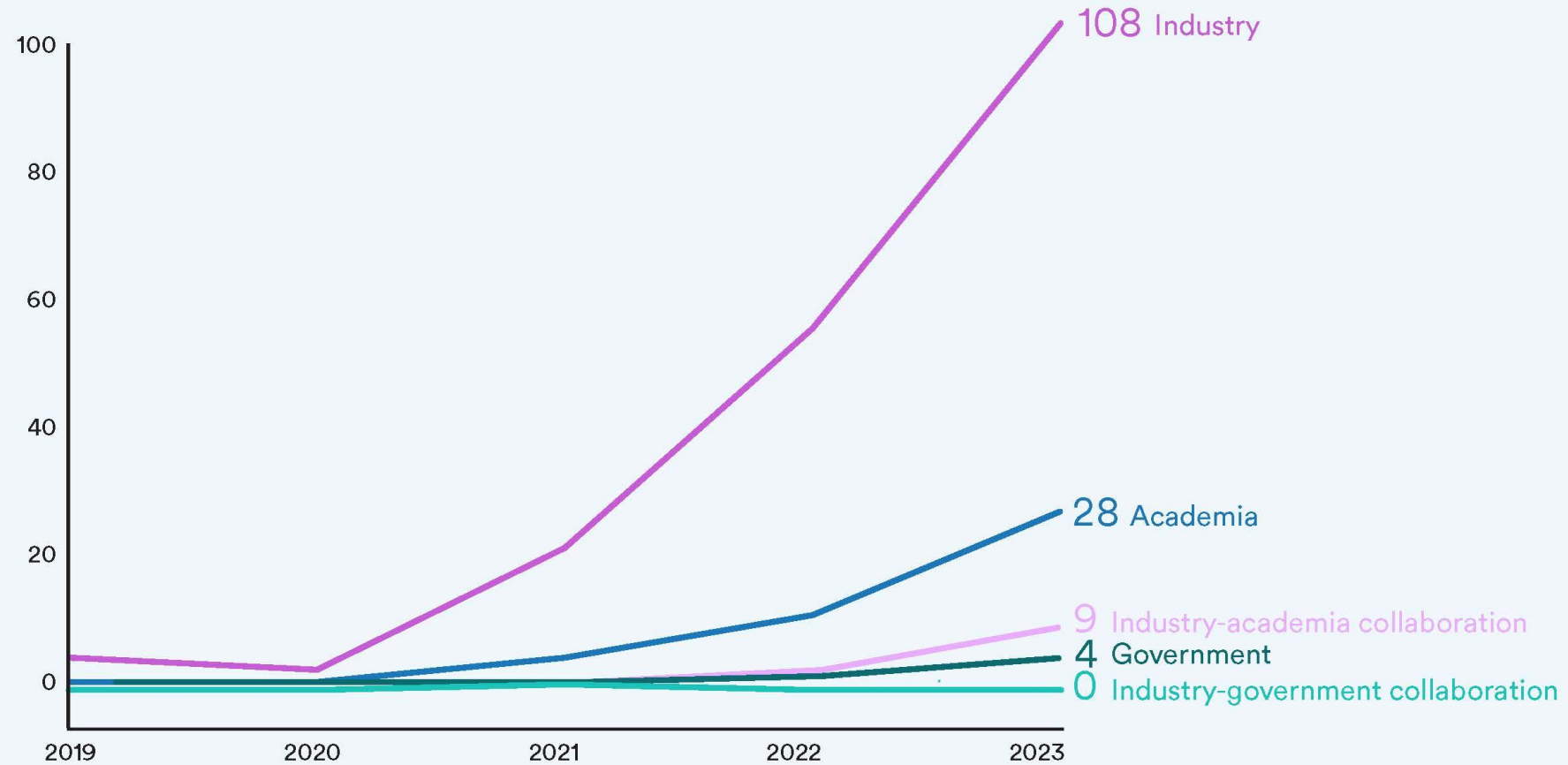


Figure 1.1.5

Number of foundation models by sector

Source: Bommasani et al., 2023 | Chart: 2024 AI Index report



OpenWeight, not Open Source...

Foundation models by access type, 2019–23

Source: Bommasani et al., 2023 | Chart: 2024 AI Index report

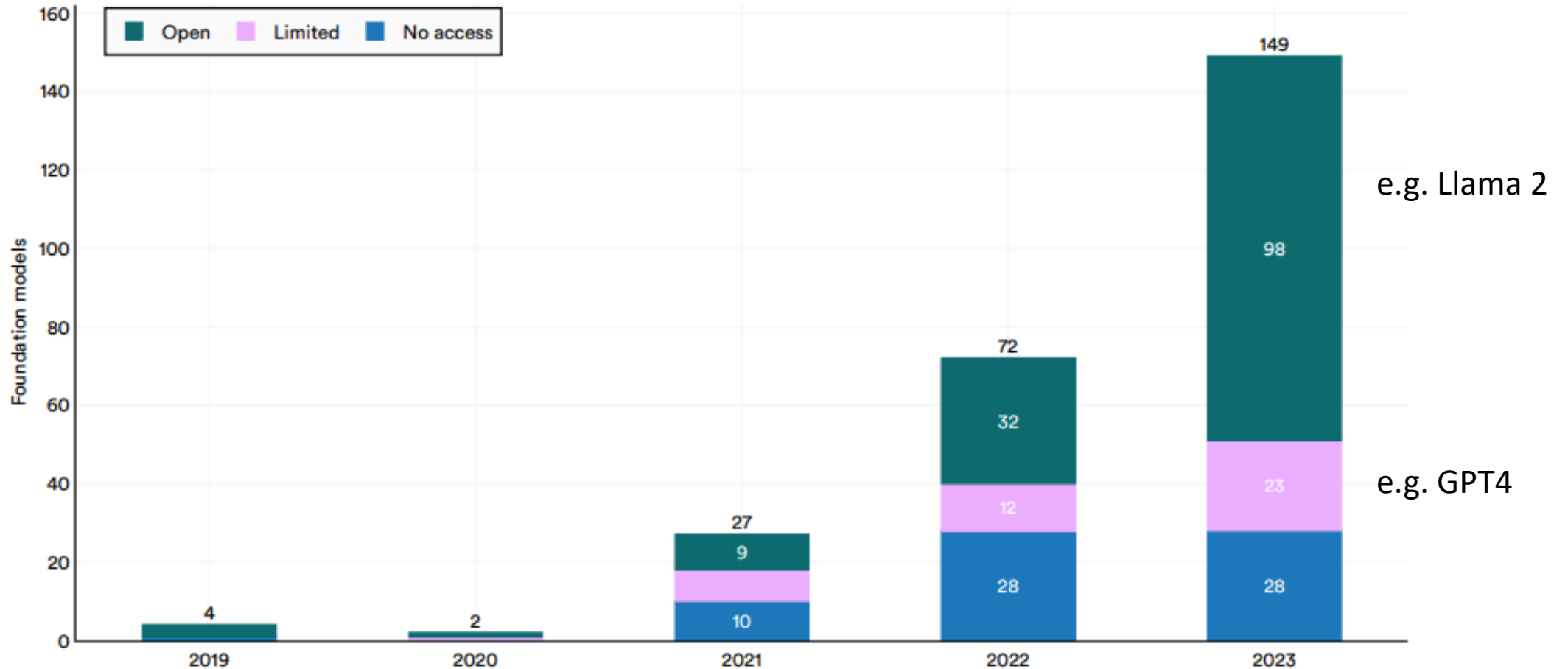
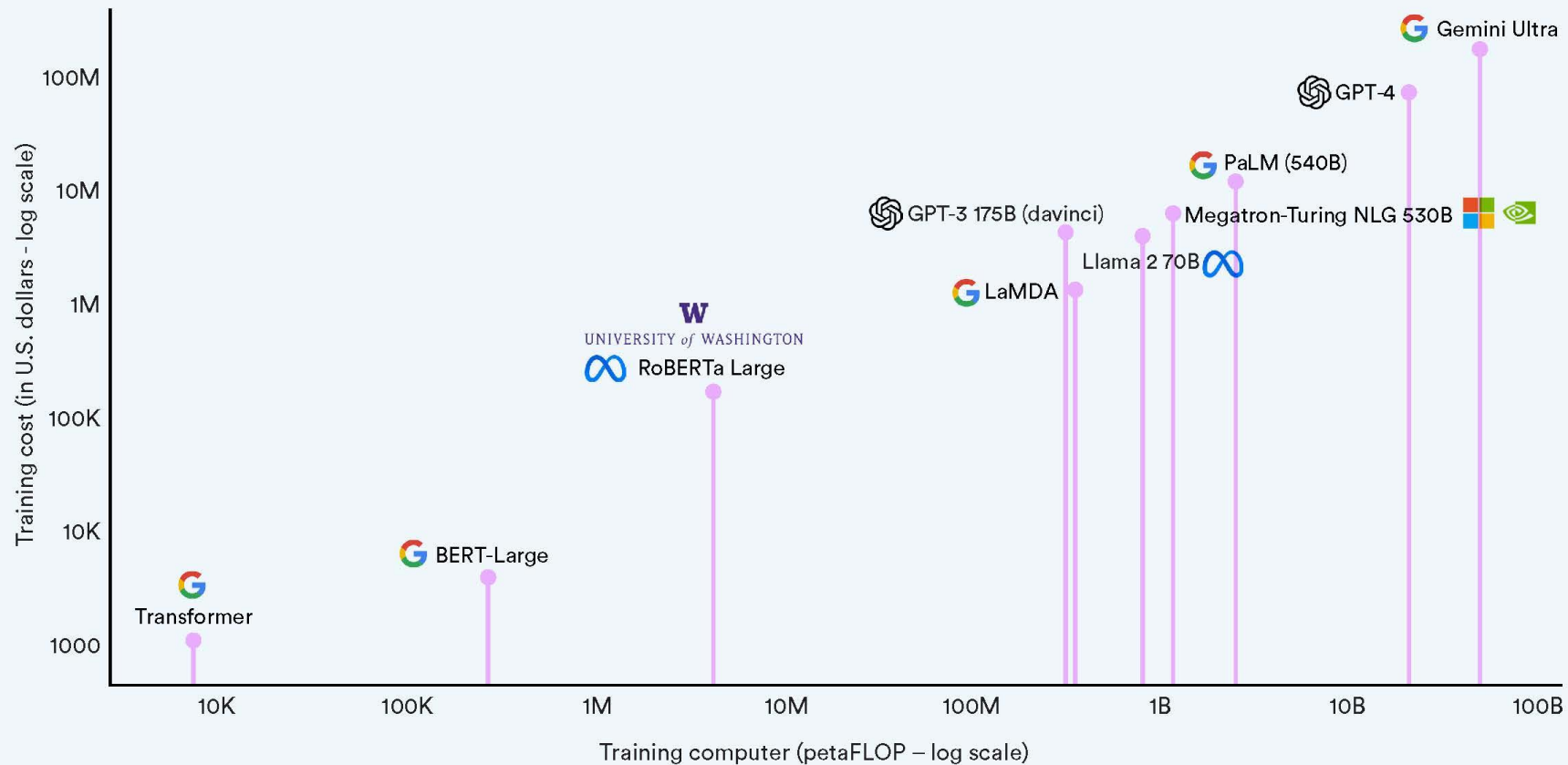


Figure 1.3.13

PETA = 1,000,000,000,000,000 Floating Point Operations Per Second

Estimated training cost and compute of select AI models

Source: Epoch, 2023 | Chart: 2024 AI Index report



Reasoning challenges / MMMU benchmark

- Abstract Reasoning (Analogy), Planning and Expert Knowledge **NO**
- Math, Visual, Commonsense and Causal Reasoning **YES**
- Moral Judgement / argumentation **?**

Massive
Multidiscipline
Multimodal
Understanding

MMMU: overall accuracy

Source: MMMU, 2023 | Chart: 2024 AI Index report

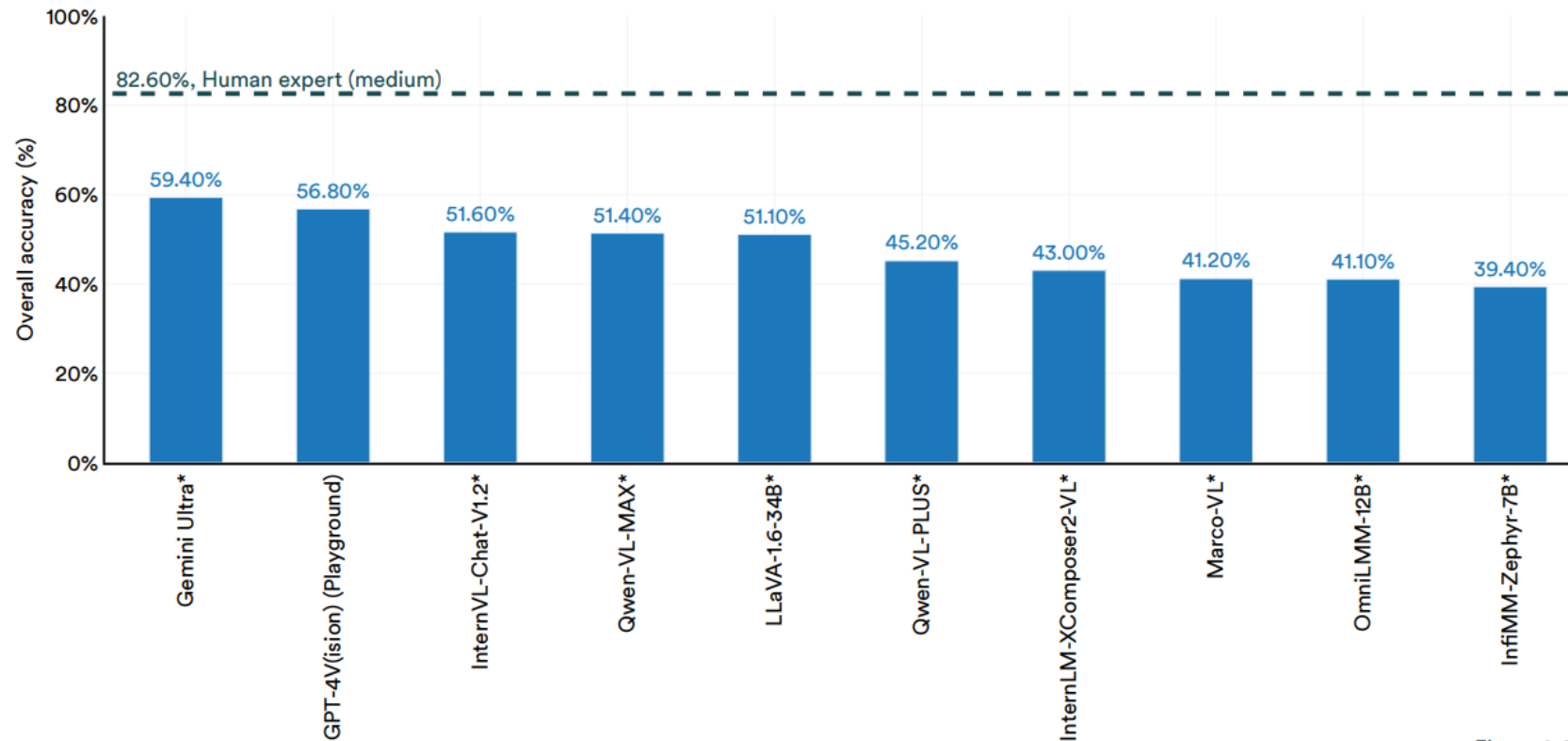


Figure 2.6.2¹²

Polluting the internet

- LLM performance over time declined on several tasks
- The curse of recursion: Training on generated data makes models forget (<https://arxiv.org/abs/2305.17493>)

A demonstration of model collapse in a VAE

Source: [Shumailov et al., 2023](#)

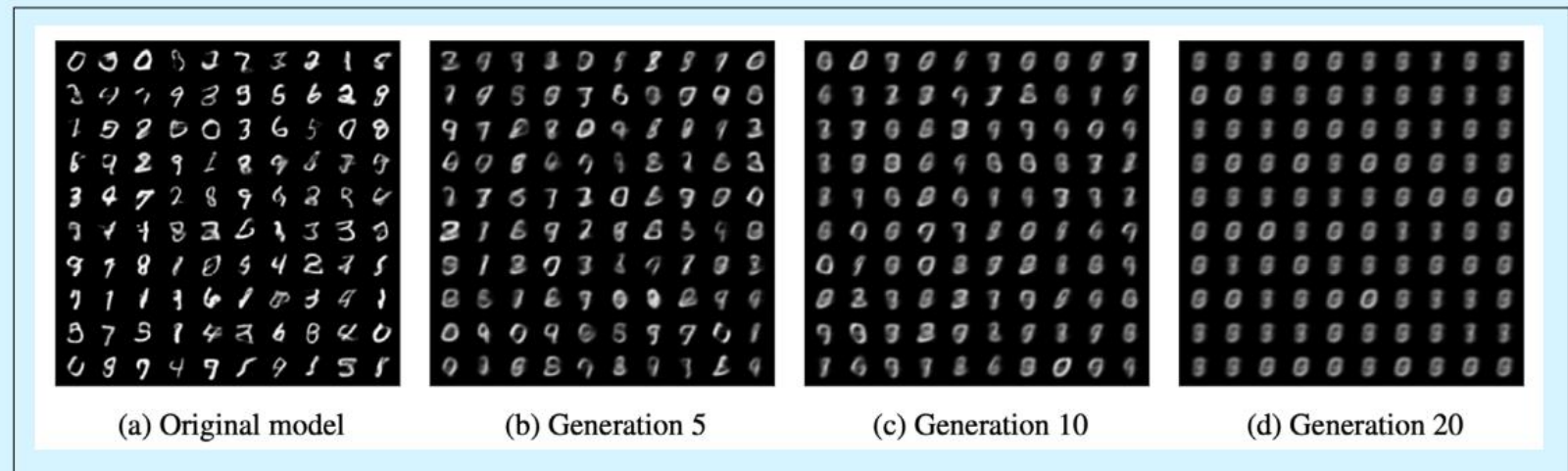
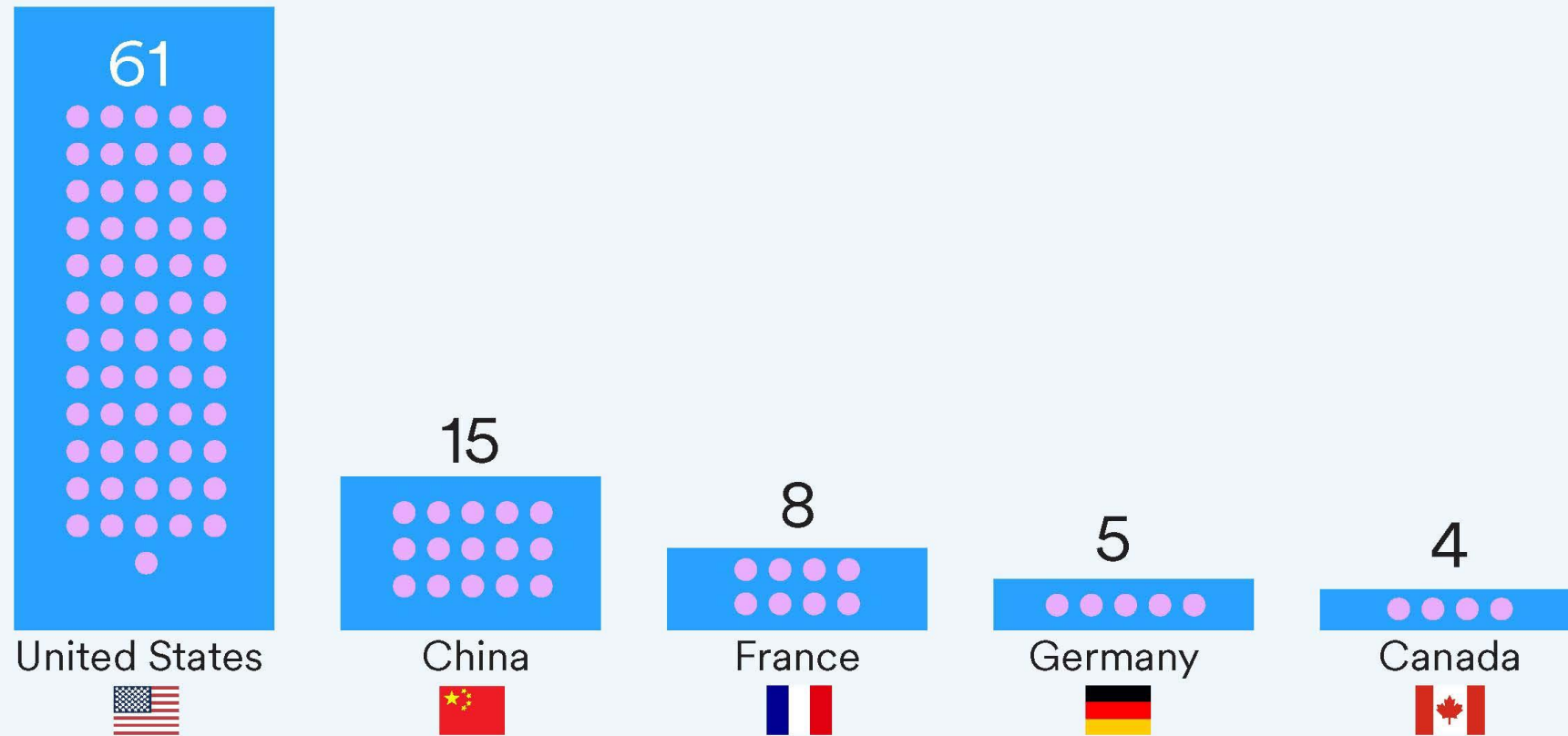


Figure 1.3.9

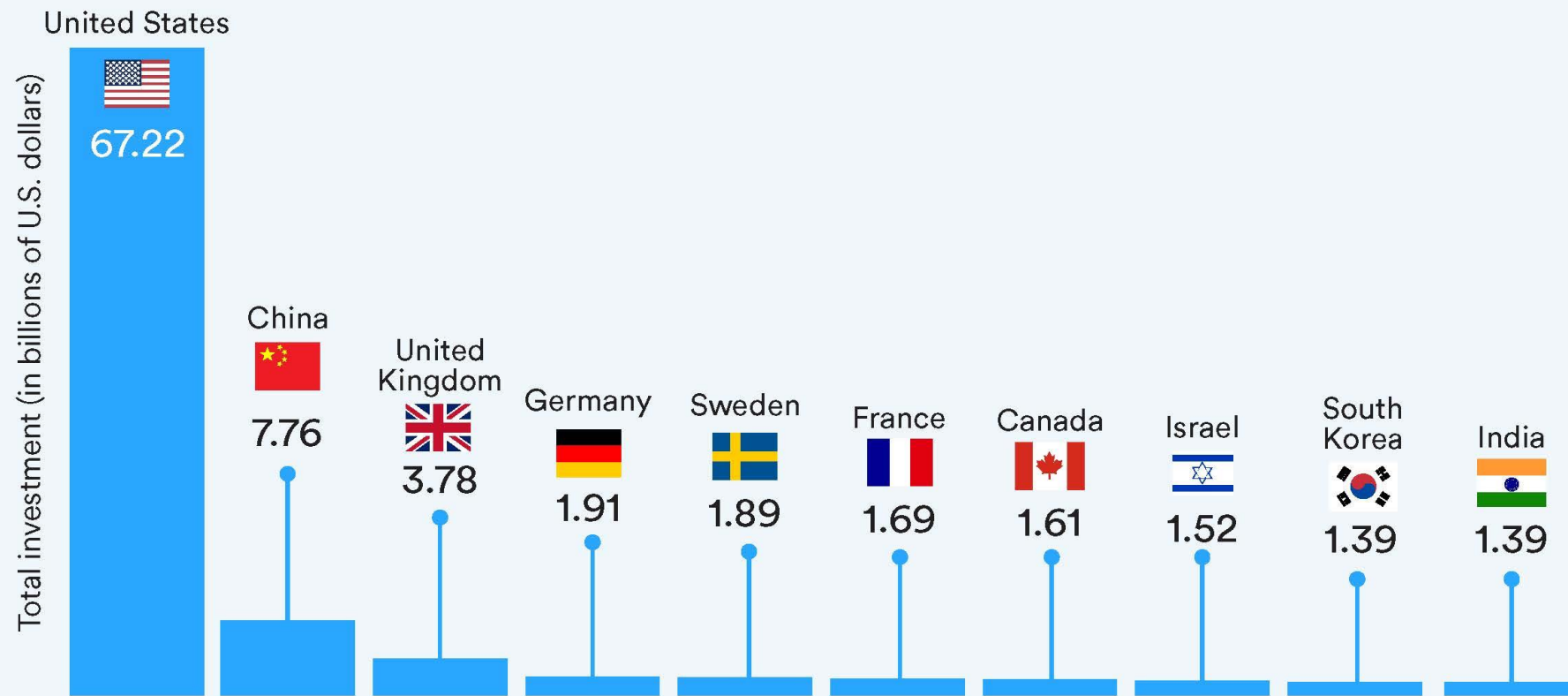
Number of notable machine learning models by country, 2023

Source: Epoch, 2023 | Chart: 2024 AI Index report



Private investment in AI by geographic area, 2023

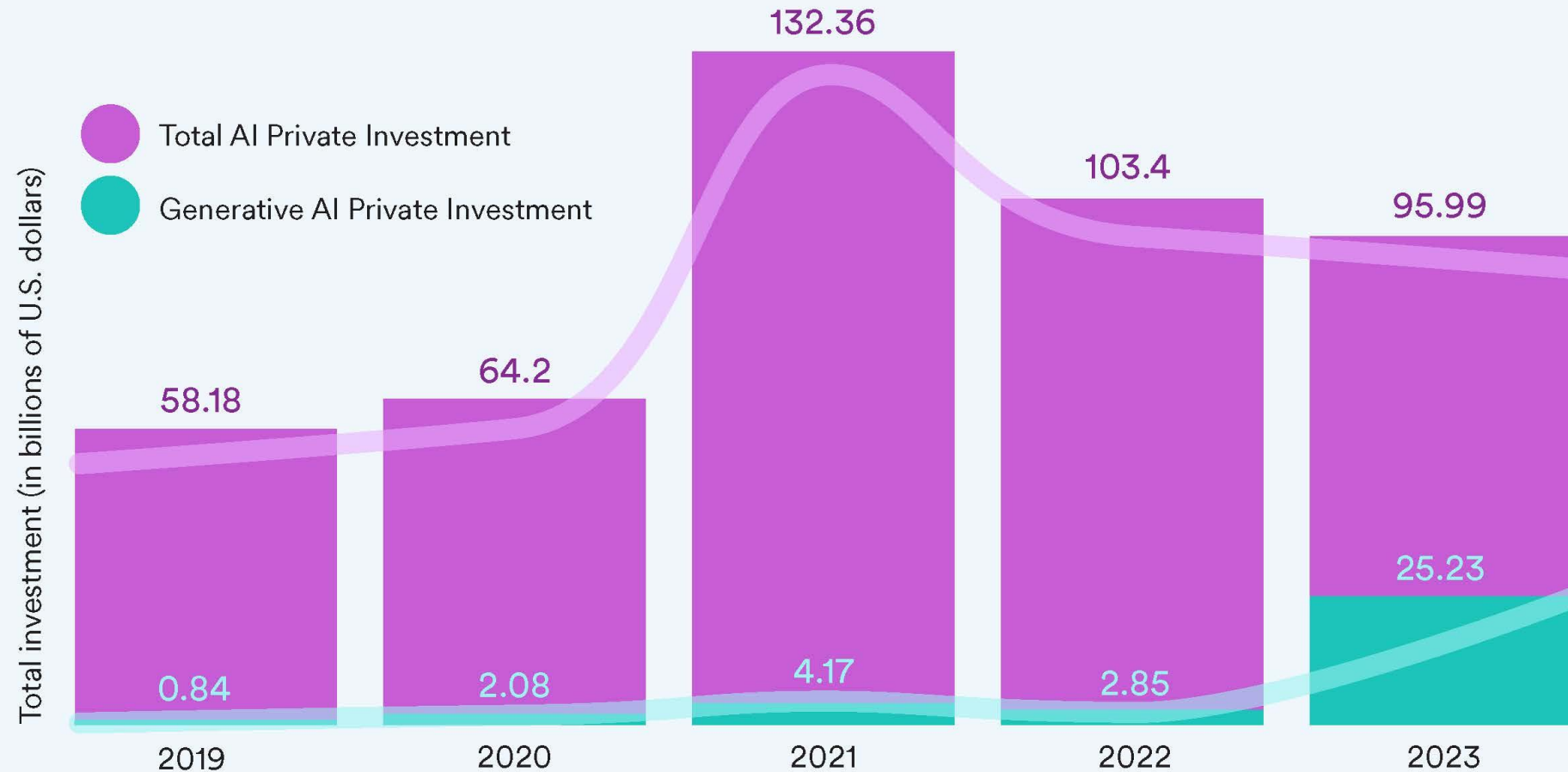
Source: Quid, 2023 | Chart: 2024 AI Index report



Private investment in generative AI

Total investment dips, but GenAI investment sees surge

Source: Quid 2023 | Chart: 2024 AI Index report



How businesses are using AI

Source: McKinsey & Company Survey, 2023 | Chart: 2024 AI Index report



Contact-center automation

26%



Personalization

23%



Customer acquisition

22%



AI-based enhancements of products

22%



Creation of new AI-based products

19%

The Main AI Innovations...

The GOOD:

- Efficient propagation of activation (message passing)
- Vector/tensor representation (embeddings)
- Attention mechanism (but attention is NOT all you need).

The BAD:

- Reinforcement Learning with Human Feedback (RLHF)
- Chain of Thought / Prompt engineering
- Scale is all you need / training on the entire internet...

Chain of Thought prompting... is a hack

- We should not try and adjust the **input** to fix a misbehaving LLM after it was trained to gobble up all the internet.
- Instead control the training: constrain the **architecture** or use the **loss function** to add knowledge – that's the neurosymbolic way!

The latest AI revolution

(GPT4, Claude, Gemini, Mistral, OLMo, etc)

- is built on Neural Networks!
- But...
 - Scale is not all you need (reasoning and planning needed as well as learning)
 - Post-hoc RLHF alignment is unethical
 - LLMs continue to hallucinate and regurgitate

Neurosymbolic AI to the rescue:

- Energy efficiency
- Safety (fact checking, self-driving cars)
- Fairness (bias vs. woke AI)

After US\$ 10B investment in the self-driving car industry (in 2021 alone)

- GM Cruise self-driving taxi SF fleet grounded following serious accident (Oct 2023)
- GM Cruise “self-driving” requires frequent help from humans working as remote assistants to get through tricky drives (Nov 2023)
- GM to resume testing its Cruise self-driving cars on public roads (Feb 2024)

There are far too many **exceptions!** The long-tail distribution problem...

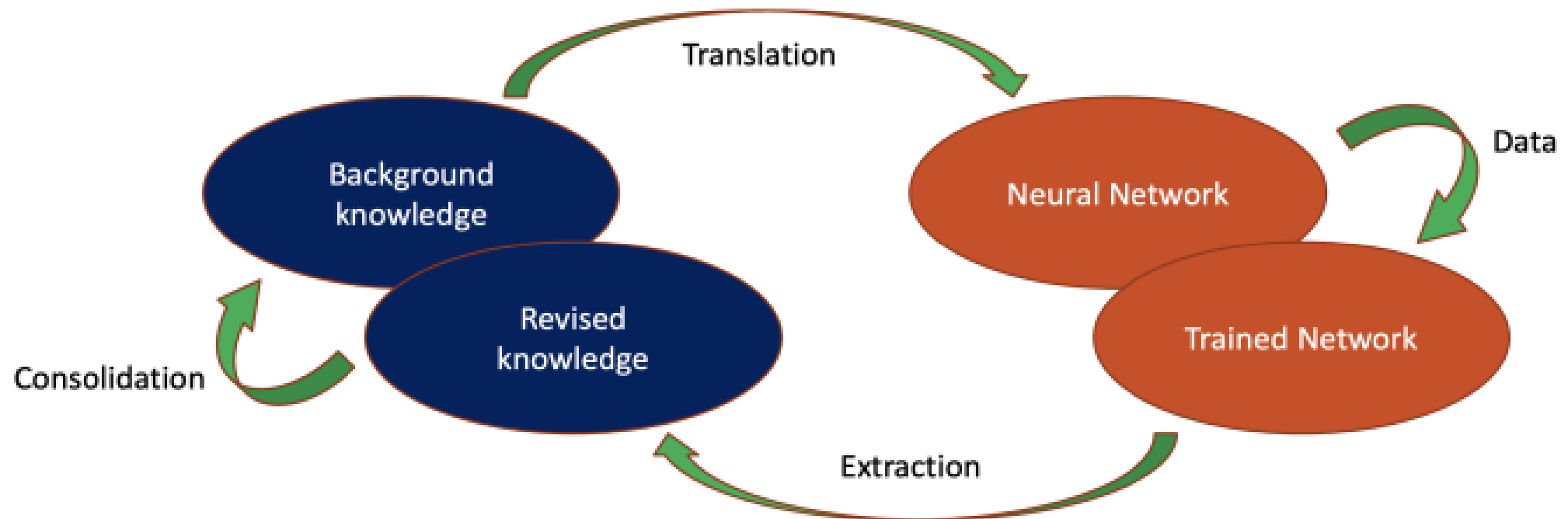
Need for learning and reasoning from data and knowledge (good old *if then else* rules alongside neural nets)

Neurosymbolic AI

You can't fact check without facts

Neurosymbolic integration: mapping logics to neural networks and vice-versa

Hybrid Systems: neural and symbolic components, e.g. LLM + KG



THE NEED FOR LEARNING AND REASONING

LEARNING AND REASONING HAS BEEN THE GOAL
OF NEURAL-SYMBOLIC INTEGRATION FROM THE
BEGINNING

CHATGPT AND OTHER ERRORS OF REASONING OF
DEEP LEARNING MAKE THE TASK URGENT

See also Valiant's keynote talk at NeSy2023:

<https://sites.google.com/view/nesy2023/home/nesy2023-recordings>

Robust Logics*

Leslie G. Valiant
Division of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138

Abstract

Suppose that we wish to learn from examples and counter-examples a criterion for recognizing whether an assembly of wooden blocks constitutes an arch. Suppose also that we have preprogrammed recognizers for various relationships e.g. *on-top-of*(x, y), *above*(x, y), etc. and believe that some quantified expression in terms of these base relationships should suffice to approximate the desired notion of an arch. How can we formulate such a relational learning problem so as to exploit the benefits that are demonstrable in propositional learning, such as attribute-efficient learning by linear separators, and error-resilient learning?

We believe that learning in a general setting that allows for multiple objects and relations in this way is paradigmatic of the more fundamental questions that need to be addressed if one is to resolve the following dilemma that arises in the design of intelligent systems: Mathematical logic is an attractive language of description because it has clear semantics and sound proof procedures. However, as a basis for large programmed systems it leads to brittleness because, in practice, consistent usage of the various predicate names throughout a system cannot be guaranteed, except in application areas such as mathematics where the viability of the axiomatic method has been demonstrated independently.

In this paper we develop the following approach to circumventing this problem. We suggest that brittleness can be overcome by using a new kind of logic in which each statement is learnable. By allowing the system to learn rules

learning and deduction are provably polynomial bounded.

1 Introduction

According to Aristotle "every belief comes either through syllogism or from induction" [1]. Computational systems that aspire to exhibit some characteristics of intelligence need to manipulate beliefs about the world. It is reasonable to ask, therefore, how useful Aristotle's dictum is for the construction of such systems. Our purpose here is to argue that the duality expressed by the dictum is fundamental. In particular, we present a formal system that encapsulates this duality and, we believe, offers a vehicle for studying the theoretical basis of such systems.

The study of syllogisms, and reasoning more generally, gave rise to mathematical logic. The mechanization of induction is the subject matter of machine learning. In this paper we suggest a formal system that encompasses learning and reasoning in an integrated way and retains the somewhat different but crucial benefits that each has to offer.

The benefits of mathematical logic that we wish to retain are the existence of a clearly defined semantics for each statement, and the existence of proof procedures that enable new statements to be derived. In particular, the well defined semantics makes possible proof procedures that are *sound*: new statements that are derived from true statements are themselves true.

The main benefits of learning that need to be retained are that it provides a mechanism by which knowledge can

A SEMANTIC FRAMEWORK FOR NEUROSYMBOLIC COMPUTING

SEMANTIC FRAMEWORK (MODEL THEORY):

STABLE STATE IN THE NEURAL NETWORK =
MODEL OF THE LOGIC THEORY

NETWORK DYNAMICS IMPLEMENTS FIXED-POINT
COMPUTATION IN PARALLEL

previous state of $h = 0$ makes $C = 0$. Updating a second time gives the state $(1, 0, 1, 1)$. This is because $A = 1$ produces $h = 1$, and the previous $h = 1$ gives $C = 1$. State $(1, 0, 1, 1)$ then produces state $(1, 0, 1, 1)$, that is a stable state equivalent to mapping (A, B, C) , starting from $(False, True, True)$, to $(True, False, True)$, which is the same result as calculating the least fixed-point of P [11]. This correspondence, once proved, makes the neural network semantically equivalent to the knowledge base. Implementing the least fixed-point operator T_P of a logic program in a neural network is a common method for encoding symbolic knowledge into neural networks [16, 8, 25] first introduced in [24].³

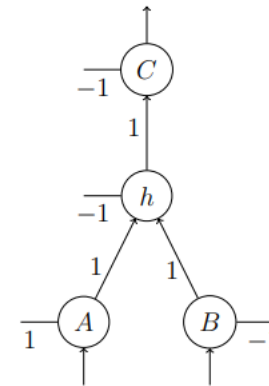


Figure 1: A simple feed-forward neural network encoding a knowledge base containing rules C if A , written $C \leftarrow A$, C if B , written $C \leftarrow B$, and fact A , written $A \leftarrow$. The parameters of the network (weights and biases) are shown next to the arrows in the diagram. With bias 1, neuron A will always produce output 1 (we say that A is *activated* in this case) for any input in $\{0, 1\}$, given the step function as activation function. With bias -1 , neuron B will output 0 for every input. Activating either A or B will always activate neuron h , since the weight (1) from either A or B to h is equal to (or larger than) the negative of the bias of h . Finally, activating h also activates C , for the same reason as above.

Another notion of equivalence that is not dependent on the fixed-point semantics of logic programming - and therefore applicable to other forms of

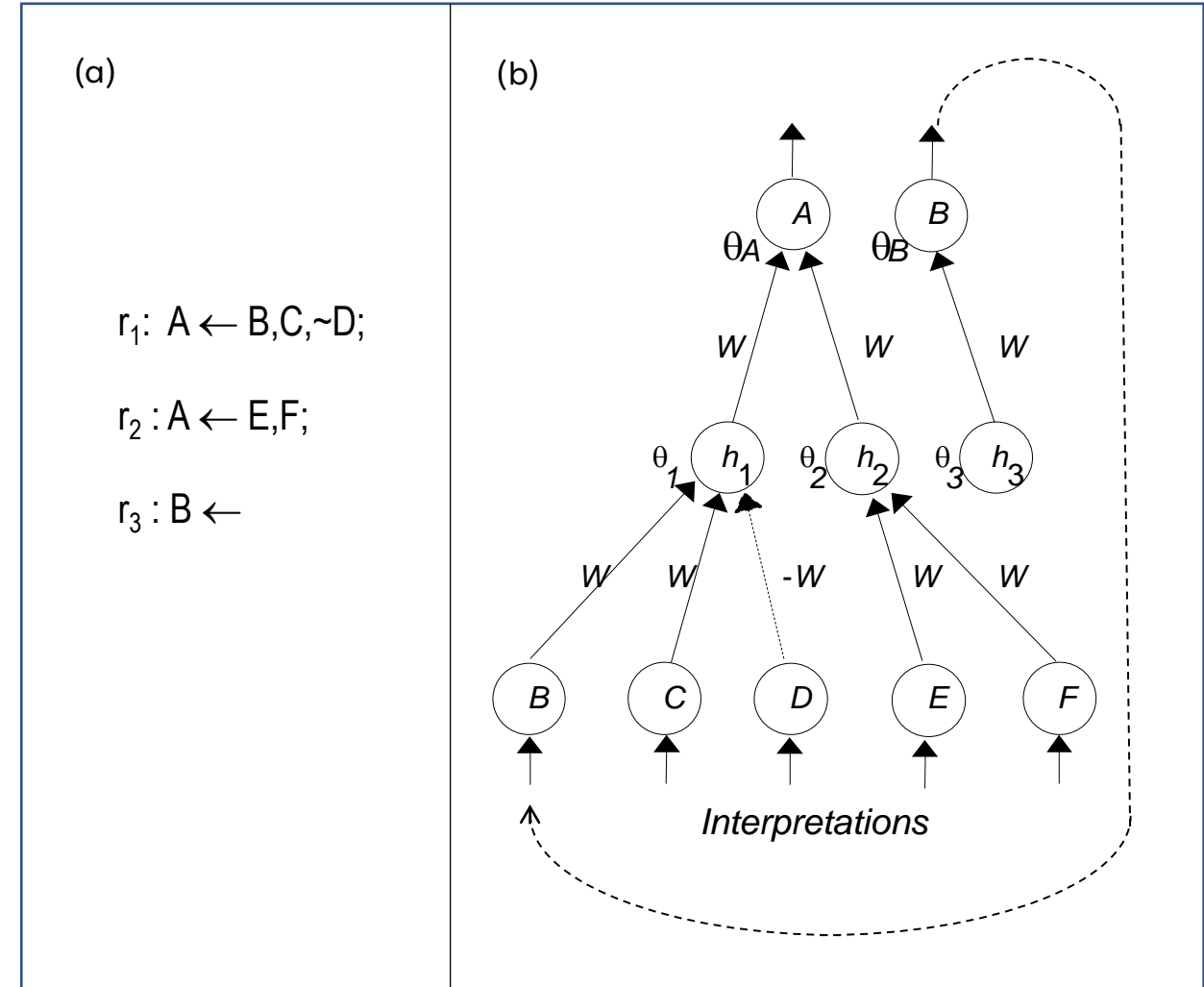
<https://arxiv.org/abs/2212.12050>

ADDING KNOWLEDGE TO THE NETWORK ARCHITECTURE

CILP SYSTEM: LEARNING IN RECURRENT NEURAL NETWORKS AND REASONING WITH PROOF OF SOUNDNESS

SOUNDNESS PRODUCES BETTER LEARNING PERFORMANCE

USE OF NEGATION MAKES CILP SYSTEM NON-MONOTONIC



CONNECTIONIST MODAL LOGIC

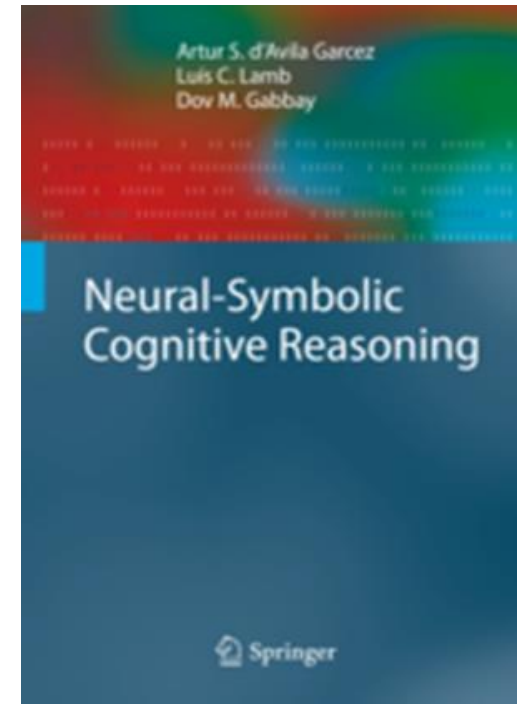
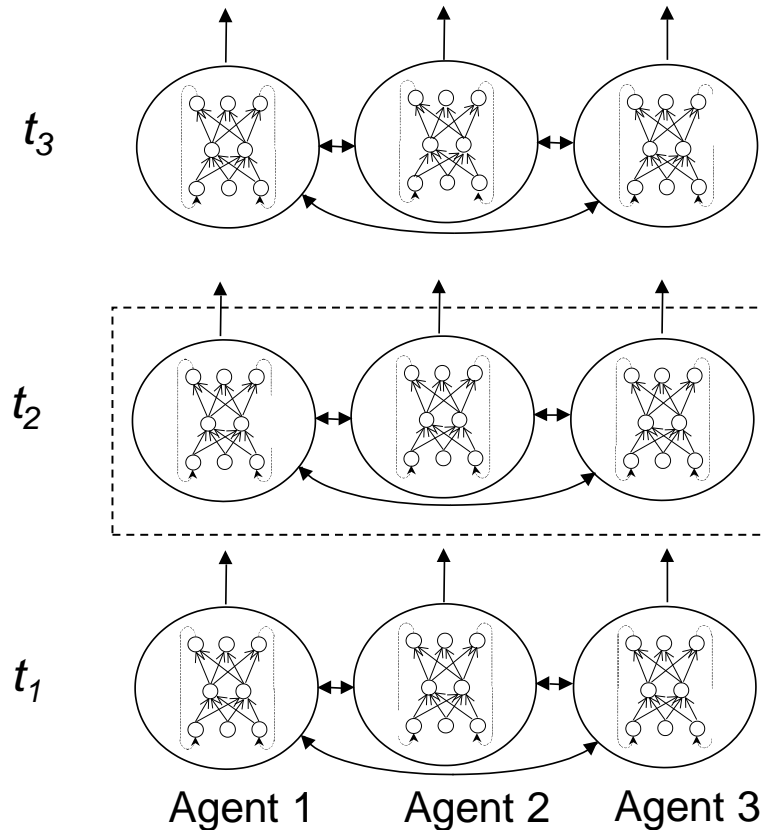
CML = Robustly Decidable
(M. Vardi)

How to get to FOL?

Neural nets are much more
complex now?

Attention layer = binary relations

SOLUTION TO THE MUDDY CHILDREN PUZZLE:

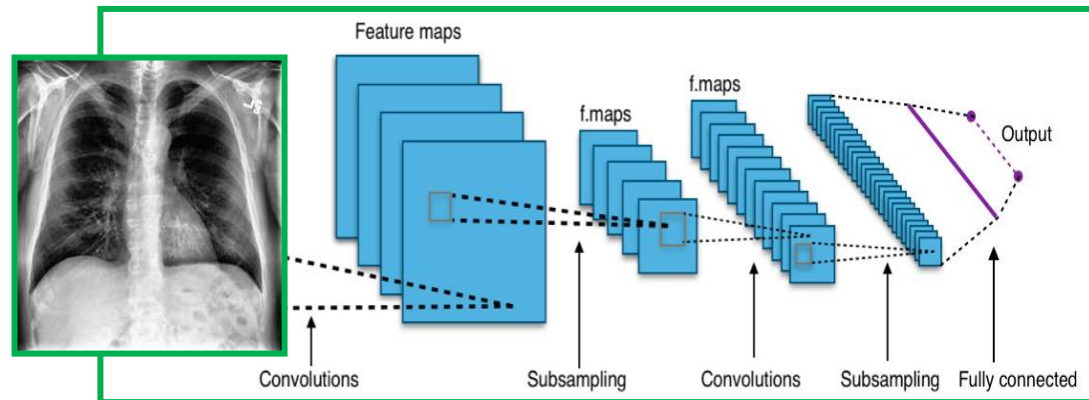


Level of abstraction

HARD/SOFT CONSTRAINTS AS LAYER ON TOP OF COMPLEX NETWORK



MEDICAL APPLICATION (COLLABORATION WITH FUJITSU RESEARCH):



ERIC
➔

-CL, -PA → pleural_effusion
CL, RW → pleural_effusion
CL, RA, -RW → pleural_effusion
-CL, PA → healthy
CL, -RA, -RW → healthy

K. Ngan, A. d'Avila Garcez, J. Townsend. Extracting **Meaningful** High-Fidelity Knowledge from Convolutional Neural Networks. IJCNN 2022.

Closing the Neurosymbolic Cycle

Domain expert co-design of AI system; expert can ask what-if questions!

Enlargement of the heart AND obscured diaphragm IMPLY pleural effusion...

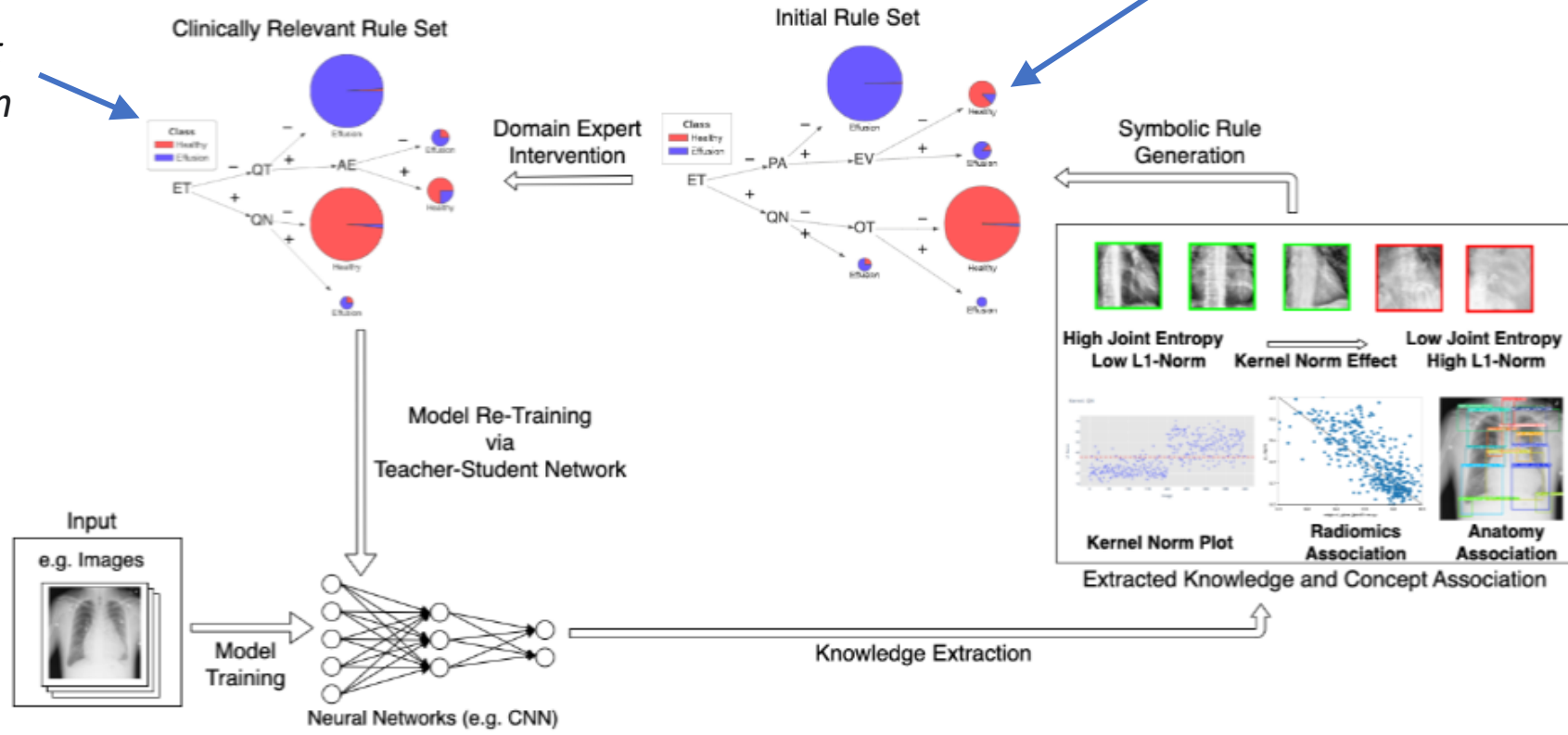


Figure 1: An overview of a neural-symbolic cycle illustrating the process of (a) extracting knowledge from a trained CNN for medical image diagnosis, (b) generating symbolic rules based on the extracted knowledge, (c) expert rule interaction and intervention to produce clinically-relevant knowledge, (d) transferring of relevant knowledge from the rules to a student CNN, closing the neuro-symbolic cycle.

Asking **what-if** questions

(knowledge extraction/XAI at the right level of abstraction)

- Extracting meaningful and compact symbolic descriptions from trained networks
- Domain expert can interrogate the system and intervene in the system
- Trustworthy AI requires knowledge extraction!
- G. Marcus's recommendation to US Senate:
 1. A safety review prior to widespread deployment;
 2. A nimble monitoring agency with authority to call things back; and
 3. Funding for research to make AI more trustworthy.

ADDING KNOWLEDGE TO THE LOSS FUNCTION

HANDLING FOL (REAL LOGIC: many-valued logic / each constant is a vector) MAPPED TO LOSS FUNCTION

LEARNING = MAXIMIZING SATISFIABILITY

REASONING IS APPROXIMATE: We measure **Reasoning Capability** instead of proving soundness

MODULARITY IS IMPORTANT PROPERTY (LTN + XAI)

Other Proof Theory approaches:

reduction to logic circuits

soft-unification

SAT solving

Example

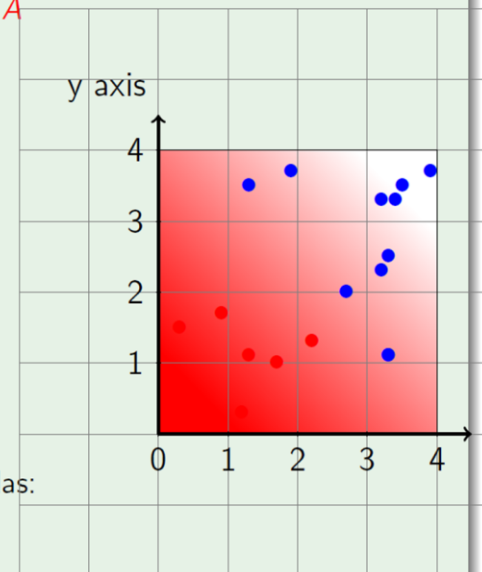
- The domain is the square $[0, 4] \times [0, 4]$;
- We have a set of examples of the class A
- And a set of examples of the class B
- We know that A and B are disjoint
- and let the shape of the membership function of the classes be

$$\sigma(w_1 \cdot x + w_2 \cdot y + w_3)$$

with $\sigma(x)$ the sigmoid function $\frac{1}{1+e^{-x}}$

- We have to find the parameters w_1^A, w_2^A, w_3^A and w_1^B, w_2^B, w_3^B that maximize the satisfiability of the formulas:

$$A(x) \wedge B(y) \wedge \forall x : A(x) \rightarrow \neg B(x)$$



Neurosymbolic AI applications

- Addressing **bias** through explainable AI (knowledge extraction): Percy et al (2021), Accountability in AI <https://arxiv.org/abs/2110.09232>
- Addressing **fairness** by adding knowledge (e.g. gender balance rule): Wagner and Garcez (2021) <https://openaccess.city.ac.uk/id/eprint/26151/>
- Neurosymbolic approach to AI **alignment**: Wagner and Garecz (2024) <https://neurosymbolic-ai-journal.com/system/files/nai-paper-729.pdf>
- Addressing **efficiency** learning more compact networks: Ngan et al (2023) Knowledge Extraction, User Intervention and Distillation <https://www.cs.ox.ac.uk/isg/conferences/tmp-proceedings/NeSy2023/paper3.pdf>

Disagreement among leading AI figures

- From LeCun's optimism

[@ylecun](#): "Engineer: I invented this new thing. I call it a ballpen! OMG, people could write horrible things with it, like misinformation, propaganda, hate speech. Ban it now!"

- To Bengio's attempt to formalize the debate around "rogue AI"

(<https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/>)

and signing open letter calling for a 6-month pause in the development of LLM (i.e. GPT5)

(<https://futureoflife.org/open-letter/pause-giant-ai-experiments/>)

This open letter was interpreted **wrongly** as “a call to stop AI research”

- To Hinton's *epiphany* about the existential risk of AI and belief that BigTech can self-regulate:

<https://www.theverge.com/2023/5/1/23706311/hinton-godfather-of-ai-threats-fears-warnings>

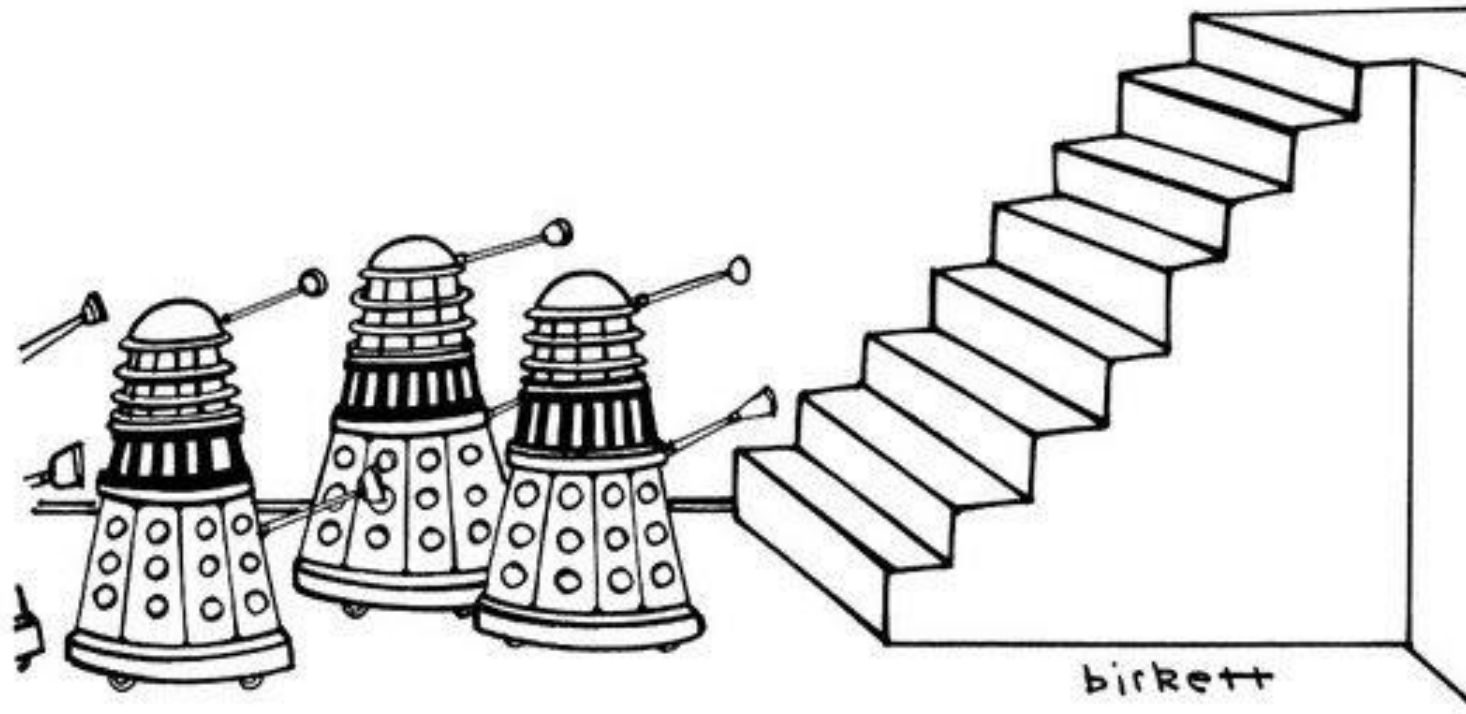
It is useful to distinguish current risk from existential risk of AGI

- We have enough problems without killer robot or fleet of self-driving cars / swarm intelligence.
- Current risk (associated with failures to regulate social media platforms; see US Senate hearing with facebook whistleblower):
<https://www.commerce.senate.gov/2021/10/protecting%20kids%20online:%20testimony%20from%20a%20facebook%20whistleblower>)
 - Bias, lack of fairness, explainability (ultimately **trust** in ML systems)
 - Misinformation at scale!

23 May 2023 (fake verified Bloomberg Twitter account): Explosion near the Pentagon causing dip in the stock market:



Embodiment matters?



"Well, this certainly buggers our plan to conquer the Universe."

AI debate 3 (Dec 2022): the main present danger is disinformation at scale
https://www.youtube.com/watch?v=JGiLz_Jx9uI

The digital poorhouse (V. Eubanks' Automating Inequality, 2018)

<https://www.nybooks.com/articles/2018/06/07/algorithms-digital-poorhouse/>

- ML-based decision making based on **biased data**, e.g. Pennsylvania's family screening tool attempting to determine who is most likely to suffer abuse at home. Data set obtained from those who seek support from council, mostly working-class families.
- Leaving aside the discussion around consciousness, sentience, embodiment, AI that "understands", qualia, intensionality...
- We're talking about partial recursive functions, Turing machines with 100B parameters.
- Emergent property? We stopped estimating **generalization** and Big Tech is not sharing their models and data, so we can't evaluate it.
- LLMs are not general, they only do one kind of thing (predict the probability of the next word/token).

In summary

- GPT4 is very impressive. It is generalist, not general; it only does one kind of thing (next token prediction).
- The way GPT4 updates is also very impressive (potential for continual learning); multimodality (images, programs, diagrams, text).
- Lack of transparency from Big Tech: how data is aggregated from the web; retraining regime?
- Embodiment or sentience not an imminent problem; people being manipulated is the main problem.
- We continue to know very little about how the (highly energy efficient) human brain works (with its 100T parameters).
- Transformers with 100B parameters are not super-Turing. Scale is not all you need. Can GPT-X become AGI? No.

What is still needed?

- Goal-directed reasoning (with querying), planning, analogy, extrapolation (FOL and higher-order logics + neural nets).
- Knowledge extraction: medical domain example / controllable network modules / smaller verifiable networks!
- Scaling up neurosymbolic programming / program synthesis and verification
- Adding (a.k.a. instilling) knowledge (e.g. facts) into neural networks in the form of soft and hard constraints.
- Logic has a major role to play: Logic is the calculus of computer science!
- Need for Accountability in AI: from principles to industry-specific accreditation, C. Percy et al., AI Communications, 2022.
- Lessons not learned from lack of regulation of social media (and impact particularly on children's mental health)

In the case of LLM specifically:

- We need to make sense of LLMs (reasoning: we don't know how it happens in LLMs; it needs to be defined even!)
- Multimodality is key / making sense of joint embeddings (image + text) more promising than text alone.
- It turns out that it is easier to learn commonsense from examples than to formalize it. We will need to be able to query and constrain LLMs with facts at the right level of abstraction, not RLHF:
neurons -> abstraction -> symbols / concepts -> neurons
- We will need nonmonotonic reasoning (jumping to conclusions), analogy, planning, extrapolation as well as online learning.
- The road to AGI: learning from fewer examples, multi-task and continual learning.

THANK YOU!

The Race is On...

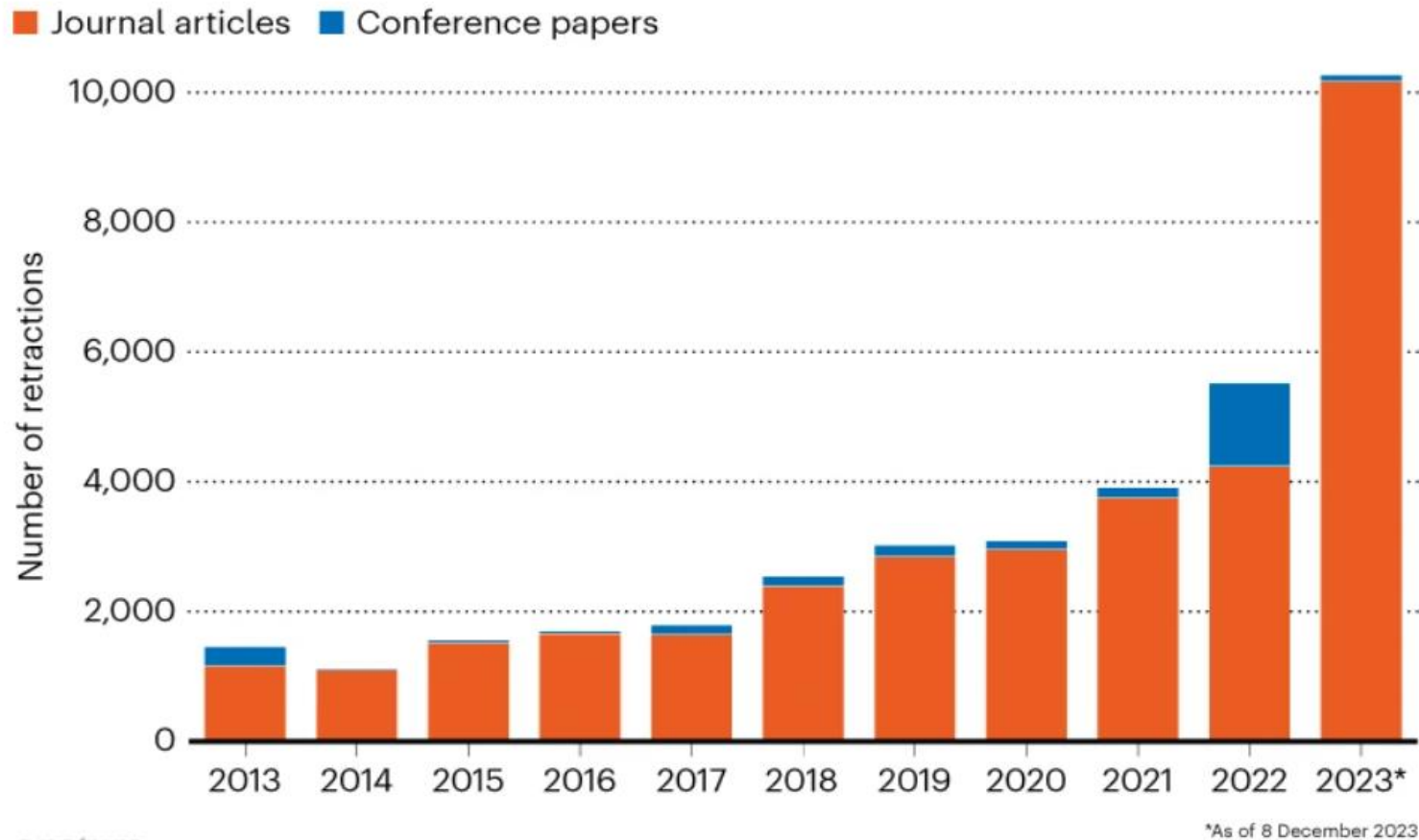
- OPENAI's Altman world tour (meeting world leaders); Sunak's bizarre softball interview with Musk...
- Ian Bremmer's TED talk The Next Superpower Isn't Who You Think:
https://www.ted.com/talks/ian_bremmer_the_next_global_superpower_isn_t_who_you_think?language=en
- Global summit on AI safety:
<https://www.gov.uk/government/news/uk-to-host-first-global-summit-on-artificial-intelligence> (with DeepMind, Anthropic, Palantir, Microsoft, Faculty.ai; **no involvement of academia or civil society?**)
- Need to control the hype around AI which is damaging AI research, e.g. DeepMind's Nature paper (June 2023) <https://www.nature.com/articles/s41586-023-06004-9> "Faster sorting algorithms discovered using deep reinforcement learning" Problem is: not a new algorithm, a better compiler optimization instead.
- Altman seeks US\$7 trillion to develop AGI while others already start to predict the start of another AI winter...
 - Despite Open AI Chaos <https://www.nytimes.com/2023/11/22/technology/how-sam-altman-returned-openai.html>
 - NY Times sues OpenAI for copyright violation; is OpenAI really a non-profit? Elon Musk sues OpenAI too...
- Geopolitics and the *chip war*: <https://www.youtube.com/watch?v=Uh4QGey2zTk>
- Nvidia benefits: In the gold rush sell shovels!

Use of GPT4 in Education and Publishing

- Not possible to fight it; embrace it!
- Consider GPT4 answers with any assignment
 - Ask students to analyse, compare and contrast, modify GPT4 answers.
 - Perhaps when no student has anything to modify we've achieved AGI...
 - Guidelines on GenAI: <https://studenthub.city.ac.uk/help-and-support/studying-online/using-ai-for-learning>
- More complicated with publishing...
- Nature's statement: no LLM tool will be accepted as a credited author on a research paper; researchers using LLM tools should document this use in the methods or acknowledgements sections.
- “An epidemic of fraud in medical research publishing”. Fabricated figures, manipulated data and flawed findings... 10k papers retracted in 2023: <https://www.ft.com/content/c88634cd-ea99-41ec-8422-b47ed2ffc45a>

10,000 research papers retracted in 2023

(ML research is based too much on conferences only;
more journal publications are needed in ML)



©nature


<https://www.nature.com/articles/d41586-023-03974-8>

Sparks of AGI?

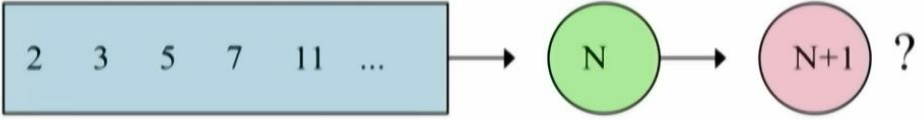
- GPT4's very impressive results (sparks of AGI paper: <https://arxiv.org/abs/2303.12712>)
 - + commonsense, affordances (use of tools), generality, programming
 - – reasoning (analogy), planning, affordances (entirely new uses)
- GPT4 gives the impression that it is pretty general; “generalist” a better word. The G in AGI is about reasoning by analogy, transfer learning, extrapolation to new situations.
 - GPT4 fails at “multi-hop” reasoning, i.e. it fails at reasoning (<https://arxiv.org/abs/2305.18654>)
 - It also fails at temporal reasoning, and therefore planning!
 - GPT4 exhibits some **reasoning capabilities** but we don't know how it does it
(Wagner and Garcez (2022) Neural-Symbolic Reasoning Under Open-World and Closed-World Assumptions: <https://ceur-ws.org/Vol-3121/paper3.pdf>)
 - GPT4's own explanations are not reliable or **measurable**.
(White and Garcez (2019) Measurable Counterfactual Local Explanations for Any Classifier: <https://arxiv.org/abs/1908.03020>)

Sparks of AGI talk (cont.)


Can you write a proof of infinitude of primes, with every line that rhymes?

 **GPT-4**

Yes, I think I can,
Though it might take a clever plan.
I'll start by noting Euclid's proof,
Which shows that primes aren't just aloof.
Assume we have a finite list,
Of primes, and that none have been missed.
Multiply them all together,
And add one, just to be clever.
The result will be a number,
That has no prime factors, oh wonder!
But every number has a prime,
That divides it, it's just a matter of time.
So we've found a contradiction,
And our finite list needs eviction.
There must be infinitely many primes,
And that's the end of my rhyming lines.



Can you draw an illustration of this proof in SVG format?



```
<svg height="300" width="500" xmlns="http://www.w3.org/2000/svg"> <!-- Draw a  
rectangle to represent the finite list of primes -->  
<rect x="50" y="50" width="200" height="50" fill="lightblue" stroke="black"/> [...]
```

This kind of apparent creativity is what seems to have impressed Hinton when testing GPT4 by asking it to **explain** an entirely new joke he had just invented. “It got the joke”!

Sparks of AGI talk (cont.)

- Programming:

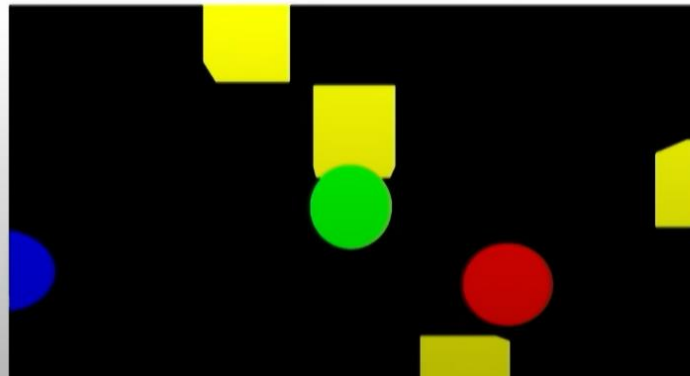
Coding with a copilot that UNDERSTANDS

Write following **3D game** in HTML with **JavaScript**:

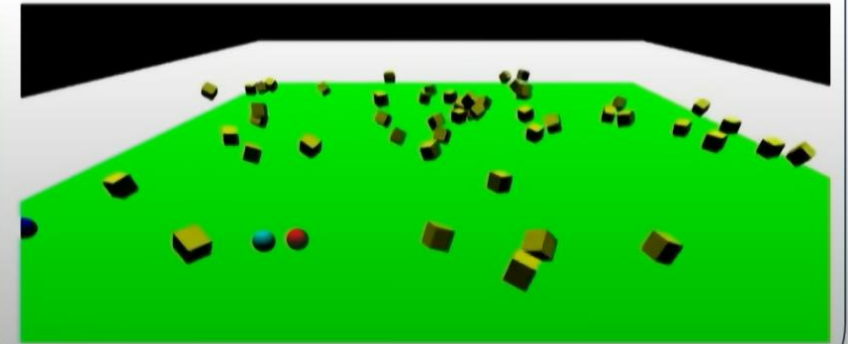


- There are three spherical avatars
- Player controls its avatar using arrow keys to move.
- Enemy avatar tries to catch the player.
- Defender avatar tries to block the enemy, but the defender sleeps for 3 seconds every 6 seconds.
- Obstacles spawn randomly and move randomly
- ...

ChatGPT



GPT-4



Expert level 500 to 1000 lines of code generated without meta-prompting, following natural language instructions precisely, even interpreting vague instructions...