Mathematics and AI

Yuri Gurevich

Computer Science & Engineering, U of Michigan

July 2025

Cybernetics or AI

"Cybernetics: Control and communication in the animal and the machine" (1948, 1961); Russian: 1958, 1968. "Artificial intelligence" was coined in 1956. At the early stage of AI, the logic approach was dominating.

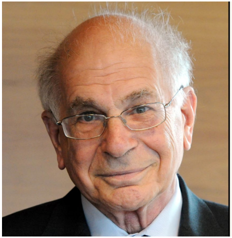
It was John McCarthy who coined the term AI, apparently to distinguish his research agenda from that of Norbert Wiener. Wiener had coined "cybernetics" to refer to his own vision of intelligent systems — a vision that was closely tied to operations research, statistics, pattern recognition, information theory, and control theory. McCarthy emphasized the ties to logic. In an interesting reversal, it is Wiener's intellectual agenda that has come to dominate in the current era, under the banner of McCarthy's terminology.

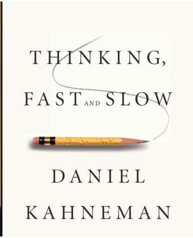
— Michael Jordan, UC Berkeley

Quotes may be edited for brevity!

Thinking fast and slow

Human thinking is driven by two distinct systems. System 1 supports *fast thinking*. System 2 supports *slow thinking*





System 1 operates automatically and quickly, with little or no effort and no sense of voluntary control. The capabilities of System 1 include innate skills that we share with other animals.

System 2 allocates attention to the effortful mental activities that demand it, including complex computations. The operations of System 2 are often associated with the subjective experience of agency, choice, and concentration.

Moravec's paradox, 1988

In the early 1970s, some of the creators of successful reasoning programs suspected that the poor performance of robotics reflected the intellectual abilities of robotics researchers. Such snobbery is not unheard of, e.g. between theorists and experimentalists in physics.

Eventually it has become clear that it is comparatively easy to make computers exhibit adult-level performance in solving problems on intelligence tests, and difficult to give them the skills of a one-year-old when it comes to perception and mobility.

Since the first multicelled animals appeared about a billion years ago, survival in fierce competition has often been awarded to the animal that could most quickly produce a correct action from inconclusive perceptions. Encoded in the large, highly evolved sensory and motor portions of the human brain is a billion years of experience.

Reasoning is effective because it is supported by this older and more powerful, though usually unconscious, sensorimotor knowledge. We are all prodigious olympians in perceptual and motor areas, so good that we make the difficult look easy.

Abstract thought is perhaps less than 100 thousand years old.

Large Language Models, LLMs

Large Language Models are content-generating deep-learning (= multi-layer) neural networks. Arguably LLMs learn like a child: first unsupervised, then supervised.

1. *Pre-training*, task agnostic, unsupervised (self-supervised) (Massive amounts of text — books, articles, websites, etc. — are collected, cleaned, and tokenized.)
2. *Fine-tuning*, task specific, supervised

<u>Godfathers of Deep Learning</u>:

Yann LeCun, Geoffrey Hinton, and Joshua Bengio shared the 2018 Turing Award for their work on deep learning. In 2024, Hinton got a Nobel Prize in Physics. LeCun was a student of Hinton.

A glimpse at internal working, and a comparison

An LLM consists primarily of: (1) a massive data file that encapsulates the knowledge learned during training and (2) an algorithm that embodies the model architecture and the probabilistic inference mechanism.

Given a query, the mechanism runs in rounds: analyze the current sequence of tokens, probabilistically infer the most likely next token, and append it to the sequence. Repeat until the sequence satisfies a specific condition.

> LLMs are trained on basically the entirety of publicly available texts on the internet, typically about 10^{13} tokens. Each token is typically two bytes, so that's $2 \cdot 10^{13}$ bytes.
>
> But developmental psychologists tell you that a four-year-old has been awake for 16,000 hours, and the amount of information that has reached the visual cortex is about 10^{15} bytes.
>
> Through sensory input we see a lot more information than we do through language.
>
> — Yann LeCun

Priors

Neither LLMs nor children start from a tabula rasa. Both have so-called “priors.” An LLM has a predefined vocabulary of tokens. Children have inborn core knowledge and predispositions due to evolution:

> Human cognition is founded, in part, on four systems for representing objects, actions, number, and space . . . Human cognition may be based, as well, on a fifth system, representing social partners.
>
> — Spelke & Kinzler, *Developmental Science* 2007

The first prior allows us to see the world split into distinct objects, the second that some of these objects are goal-pursuing agents. The other two priors are related to counting and primitive geometry respectively.

Are LLMs intelligent?

Many people say No. For example, here's Edward Gibson, MIT psycholinguistics professor & the head of MIT Language Lab:

> LLMs are doing the form. They're doing it really, really well. But are they doing the meaning? No, probably not. They really don't understand the meaning of what's going on.

I disagree. Quantity turns into quality. (Количество переходит в качество.) Already in 1997, playing against Kasparov, IBM's Deep Blue made move that looked positional, subtle, and strategic. Kasparov thought that a human grandmaster might be assisting Deep Blue.

In 2015, DeepMind's AlphaGo defeated a GO champion. One particularly famous move was unexpected, highly unconventional, even thought to be a mistake at first.

Are LLMs intelligent?

Hinton says Yes.

> LLMs turn words into features and make these features interact. From those interactions they predict the features of the next word. These millions of features and billions of interactions between features that they learn, are understanding.
>
> This is the best model we have of how we understand. So it's not like there's this weird way of understanding that these AI systems are doing and then this is how the brain does it. The best that we have, of how the brain does it, is by assigning features to words and having features and interactions.

LLMs hallucinate, admits Hinton, but so do we.

> Anybody who's studied memory knows that people are just like these large language models. They invent stuff and, for us, there's no hard line between a true memory and a false memory.

Are LLMs intelligent?

Ilya Sutskever says Yes, absolutely. Russian speaking (born in Gorky), a student of Hinton, he was the Chief Scientist at Open AI until 2024 when he co-founded Safe Superintelligence (Palo Alto, Tel Aviv).

If you value intelligence above all other human qualities, you're gonna have a bad time.

LeCun says Not yet.

They are missing essential components: understanding the physical world, persistent memory, reasoning, and planning.

Comment. To say the obvious, intelligence comes in different forms and degrees. Think of dogs, dolphins, octopuses . . . LLMs are intelligent in a sense, to a degree. Artificial intelligence is bound to be incomparable to ours; it is not optimized for survival. Airplanes don't flap wings but fly nonetheless.

AI, Reasoning, and Logic

AI badly needs reasoning, and currently reasoning is a hot issue in AI. Consider LeCun's "essential components."

1. Understanding the physical world involves reasoning — from evading predators (fast thinking) to quantum physics (slow thinking).
2. Persistent memory will help with maintaining coherence and consistency. In humans, persistent memory is vital for incubation of ideas, for cross-pollination of ideas from different domains, and for reasoning by analogy.
3. Reasoning
4. Planning surely involves reasoning.

To me, logic is the science of reasoning, any kind of reasoning. Hence AI needs logic.

A cautionary tale of infinitesimals

With the birth of calculus, the use of infinitesimals became prominent, even though the conceptual foundations were murky.

$$\frac{d(x^2)}{dx} = \frac{(x+dx)^2 - x^2}{dx} = \frac{2xdx + (dx)^2}{dx} = 2x + dx \stackrel{?}{=} 2x$$

Logicians didn't care, still doing Aristotelian syllogistic logic. (If there were mathematicians at the time who did not pay attention to calculus, we don't know their names.)

The crises of rigor led to the arithmetization of analysis (reduction to integers & logic), culminating in the ε, δ approach.

It was another crisis in the foundations of mathematics (spurred by set theoretic paradoxes) that led to a flourishing of mathematical logic and eventually to a rigorous logic of infinitesimals, *nonstandard analysis*, developed by Abraham Robinson in the 1960s. But it was too late to replace the ε, δ approach as a standard foundation for the calculus.

Few-shot learning, 2020

It was observed in an OpenAI paper (Brown et al. in arXiv) introducing GPT-3 (175 billion parameters), that LLMs can learn from examples in a prompt, without changing parameters!

Example

Translate English to Russian:
Hello → Привет
Thank you → Спасибо
How are you? → Как дела?
Good morning → ?

Prompt Engineering emerged. Researchers and practitioners experimented with how to craft prompts to get models to perform desired tasks.
Prompt design became an informal programming language of sorts for LLMs.

Chain-of-thought prompting, 2022

Discovered in a Google Research paper (Wei et al. in arXiv).

Prompt:

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Reply:

A: The answer is 11.

Prompt:

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Reply:

A: The answer is 27. X

Prompt:

Q: (Same as it is on the left)
A: Roger started with 5 balls. 2 cans of 3 tennis balls each, is 6 tennis balls. $5 + 6 = 11$. The answer is 11.
Q: (Same as it is on the left)

Reply:

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. V

LLMs are competitive with strong undergrads

Here's one illustration. MATH Dataset (Math Aptitude Test of Heuristics), introduced by Hendrycks et al. (UC Berkeley) in 2021, consists of 12,500+ competition-style math problems from high school and early college Olympiads. Problems span algebra, geometry, counting & probability, number theory, and pre-calculus.

GPT-3: 5% accuracy
GPT-3.5: 30%
GPT-4: 40–50%
GPT-4 with chain-of-thought and "self-consistency" (sample multiple reasoning paths and choose the most frequent answer): 55–60%

Limitations of current AI:

- Fragility. A plausible-looking proof may gloss over crucial steps, may be false.
- Shallow abstraction. No real insight. No self-reflection.

Is there a ceiling blocking the progress?

Nobody knows.

I am just an observer, an arm-chair general. But everybody is entitled to have an opinion. Here's mine, for whatever it's worth:

The current progress is based on fine-tuning. I doubt that it will turn LLMs into good mathematicians. Special math pre-training may be more effective. It would not be cheap. It would require:

- An expansive pre-training corpus that includes numerous examples, counterexamples, metaphors, dialog, chalkboard talk, problem-solving sessions, etc.
  The intention is to give LLMs a child-like immersion.
- Learning from playing with math, self-supervised theorem discovery (the Moore method).
- Encouraging LLMs to conjecture, analogize, etc.

What is mathematics, anyway?

What is mathematics?
A tool.

How does the tool work?
Idealized worlds are created and explored. Sometimes a real-world problem can be solved if the situation can be approximated by such an idealized world.

Example.
Suppose that I take care of your sheep for the winter. When I return, you want to verify that you have all the sheep. You can count your sheep by pointing at them, one at a time, in some order. This involves the ideal world of natural numbers and applies to many other objects but not to everything. It also involves a theorem (maybe our first theorem) that the count does not depend on the order in which you count the sheep

How do we learn math? 4 obvious remarks

1. This is not obvious. The 1954 "congruent triangles" story.

2. While deduction is indispensable in math, it is secondary. "It is by logic that we prove," wrote Poincare, "but by intuition that we discover."

3. Think of animals trying to find their way in a dark room. Neurophysiologists found that a map of the room is created in the brain. Mental maps may be a foundation of our ability to create mental landscapes of complex mathematical worlds.

4. We don't learn math the same way. Here are 2 examples.

> Vladimir Arnold contrasted Israel Gelfand and Gelfand's mentor, Andrei Kolmogorov: "Imagine they both arrived in a mountainous country. Kolmogorov would immediately try to climb the highest mountain. Gelfand would immediately start to build roads."
>
> — T. Maugh, LA Times, Sept 16, 2014

Saharon Shelah

Summary

(1) Will AI become better at math than an average mathematician?

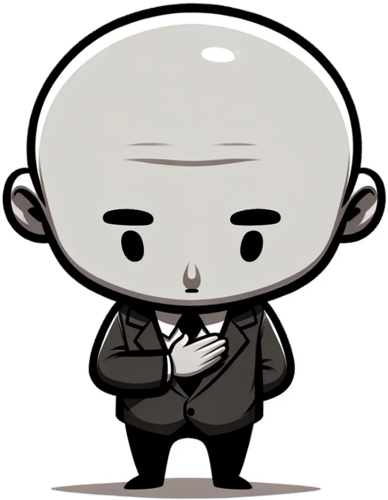
(2) If yes, then when?

My answers:

(1) Yes, inevitably.

(2) This is a matter of cost/benefit analysis. Now there are arguably more cost-effective applications of AI, like robotics.

Thanks!



Mathematical tea in St. Petersburg