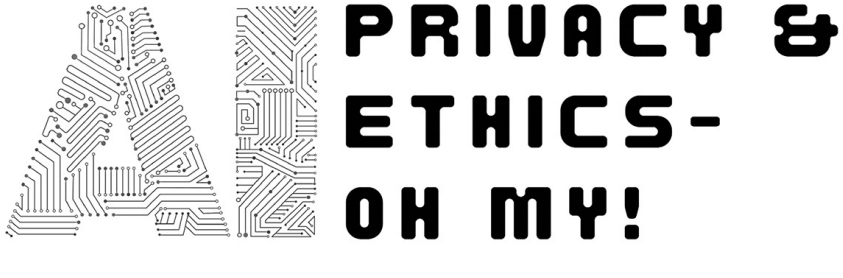
After 75 Years of AI, Can Machines Think?

Tim Finin

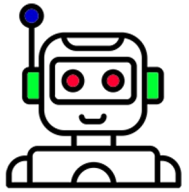
April 1, 2024



Today's Talk

- Some AI history
- What's led to recent advances in AI
- How generative AI created an inflection point
- What GAI systems can and can't do well yet
- Risks and potential misuse of AI technology
- What's next
- Can machines think?

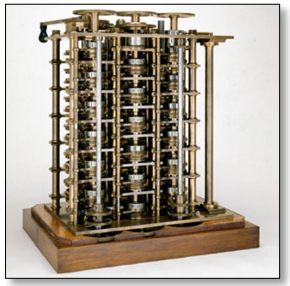
BLUF: My perspective based on 50 years in AI



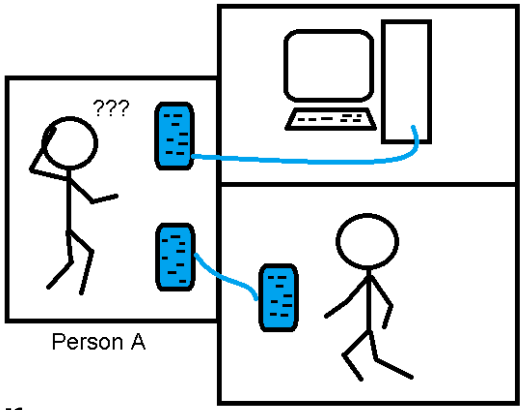
- We've not solved all of AI's problems nor yet developed what some call an **AGI** (Artificial General Intelligence)
- **ChatGPT** and similar systems (e.g., **Copilot**, **Gemini**, **Claude**) show remarkable and useful capabilities that
  - Are being integrated into software systems like web browsers, editors, programming environments, spreadsheets, and more
  - Can and will be improved by adding current and future AI advances
- Their **impact on society** will be like that of the **Web**, which was introduced about 30 years ago
- **Amara's Law**: "We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run"

Charles Babbage & Ada Lovelace

- **Babbage** designed the first computer in 1832 and saw it as just a number cruncher
- Colleague **Lovelace,** the first programmer, saw that numbers can represent other entities, enabling machines to reason about anything!
- But she wrote: "*The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform*"



Turing Test, 1950



- Polymath **Alan Turing** devised a simple test to evaluate if a computer can think in 1950
- There are three rooms with a person, computer, and an interrogator
- The interrogator communicates with others via a **teleprinter** and tries to distinguish the person and machine
- The person acts naturally, and the machine tries to fool the interrogator into believing it is the person
- If the machine succeeds, we conclude it can think

Turing, Alan, Computing Machinery and Intelligence, Mind, v596, n256, pp.433-460, 1950

1956 Dartmouth AI Project

AI pioneer John McCarthy introduced the term artificial intelligence in a proposal for a 1956 summer workshop at Dartmouth

Five of the attendees reunited in **2006**

We propose that a **2-month**, **10-man** study of **artificial intelligence** be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of **learning or any other feature of intelligence** can, in principle, be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use **language**, form **abstractions** and **concepts**, **solve kinds of problems** now reserved for humans, and improve themselves. We think that **a significant advance** can be made in one or more of these problems if a carefully selected group of scientists **work on it together for a summer**.

What is AI?

Q. What is **artificial intelligence**?

A. It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable.

John McCarthy

http://www-formal.stanford.edu/jmc/whatisai/

Ok, so what is intelligence?

Q. Yes, but what is intelligence?

A. **Intelligence** is the computational part of the ability to **achieve goals** in the world. Varying kinds and degrees of intelligence occur in people, many animals, and some machines.

John McCarthy, again

http://www-formal.stanford.edu/jmc/whatisai/

Eliza appeared Intelligent in 1966

- **Eliza** was a very simple chatbot many believed understood human language
- Its default model simulated a therapist
- **Weizenbaum**'s goal was to show how easy it is for a program to **appear** intelligent
- **Eliza effect**: a term used for ascribing understanding to systems with good text interfaces

…but wasn't ☹

AI Summers and Winters

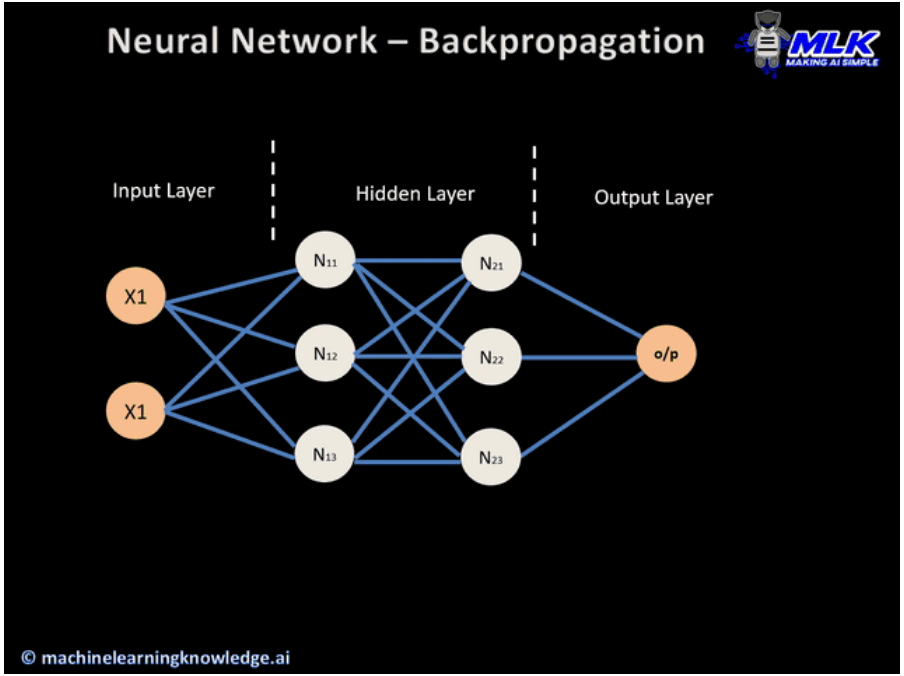
- The AI field has had its ups and downs
  - 50-60▲, 70s▼, 80s▲, 90s▼, 00s▲ then ▼, 10s ▲, 20▲▲
  - Like the stock market, the overall trend is up
- Recent advances have come from new ideas, especially from **machine learning**
- These are supported by increased computing power (e.g., parallel computing) and massive training data availability (e.g., the Web)

We're at another AI inflection point

- OpenAI released ChatGPT in late 2022 showing the potential of Generative AI (GAI) systems
- ChatGPT conversed with people to answer questions, generate text, and generate images using DALL·E
- Google, META, Apple, and others have since released similar systems, & open-source ones are also available
- Causing many to see their **benefits,** triggering interest, excitement, and some hype, but also
- See their **shortcomings** and **risks,** raising concerns

Neural Network Technology

- GAI systems use huge neural networks with billions of nodes and a trillion+ parameters
- Trained with data for tasks like text and image understanding
- Advances in the 1990s and 2000s (e.g., **backpropagation**) enabled them to be effective
- High-performance, parallel computing systems are needed to train them



Neurons & links between them have weights that are slowly learned to support some tasks using training data

Large Language Model (LLM) networks

- **Language models** have been used for decades to support many applications, e.g., autocomplete, speech recognition, translation
- **Large Language Models** (LLMs) like OpenAI's **GPT-4** use **neural networks** trained on huge text corpora from the Web and books
- Given previous words $w_1 \ldots w_{N-1}$ as context, predict the next words

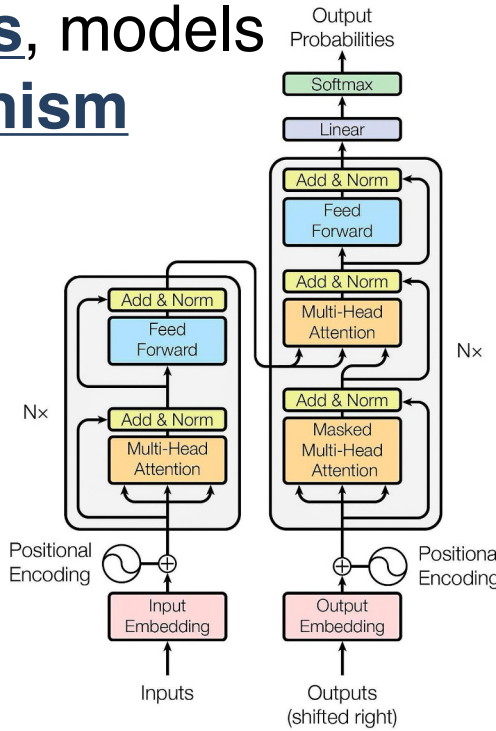
  $p(w_N = v | w_1 \ldots w_{N-1})$, e.g., $p(w_N = \text{meowed} | \text{The, fluffy, cat}) = 0.002$
- Some call them **stochastic parrots**, which can generate coherent word sequences but don't understand what they say

Stochastic Parrot?

What is a Transformer LLM?

- LLMs in current GAI systems use **Transformers**, models with **word embeddings** & an **attention mechanism**
  - These features make them much more powerful
- The model & training corpora sizes make them **expensive to create,** in cost and energy

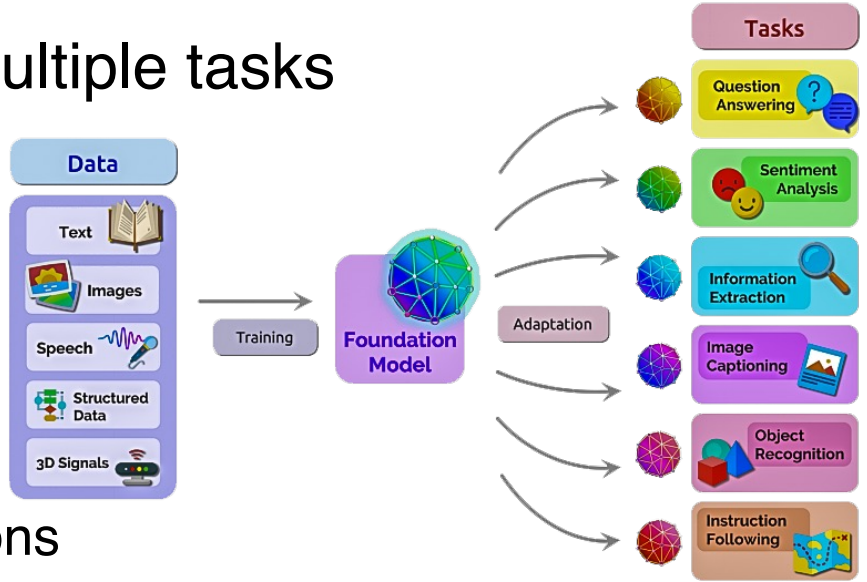
  GPT-4 has ~1.8 trillion parameters over 120 layers and cost more than $100M to train
- Smaller pretrained LLMs are available for researchers, like Meta's **LLaMA** and **Hugging Face** models



From Attention Is All You Need

Ok, I have a LLM, now what?

- LLMs are called **foundation models** since they are the basis for building or supporting **multiple AI applications**
  E.g., language translation, sentiment detection, summary generation, question answering, coding assistants, and more
- LLMs are **fine-tuned** to support multiple tasks
  - Extends neural network with layers for an application type, e.g., summarization
  - Use supervised learning to train result with sample inputs and desired outputs
  - Then, use reinforcement learning via human feedback to improve applications

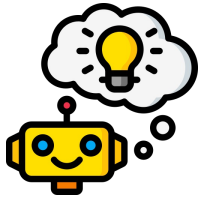


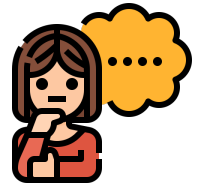
Problems with current GAI systems

1. LLMs are mostly unable to cite sources for confirmation
2. They can "hallucinate" some facts
   Q: When did Leonardo da Vinci paint the Mona Lisa?
   A: Leonardo da Vinci painted the Mona Lisa in 1815.
3. They lack common sense reasoning
   25 US states have a town named Washington, but there are also only 9 US towns named Washington
4. They have poor mathematical and logical reasoning
5. Can learn social bias & misinformation from training data
6. Can be *poisoned* by ingesting intentional disinformation

Partial list of problems and errors. They're being addressed and some are partly solved

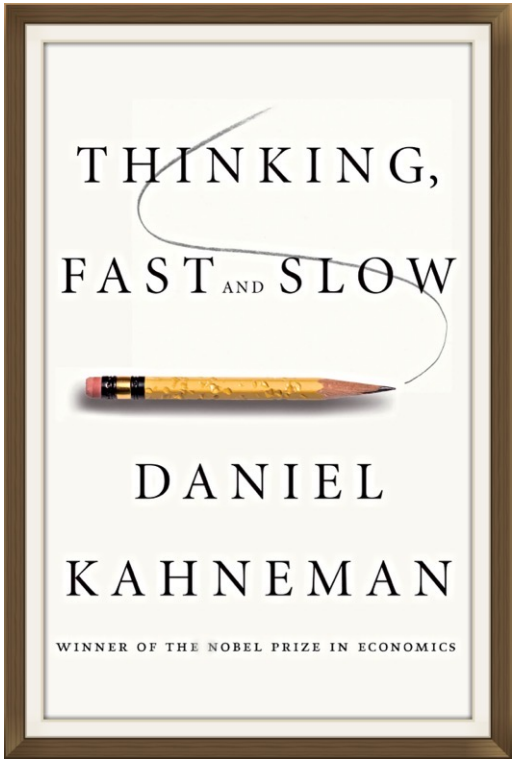
Reasoning by People and Machines





- AI up to the 1990s, AI reasoning was mostly done with programs based on logic and formal algorithms
- People can reason, but not always well, e.g., negation and disjunction are particularly difficult
- People use many kinds of reasoning strategies, most of which are neither *sound* nor *complete*
- But people do learn to reason more formally, using logic, math, probability, and other systems

Thinking Fast and Slow



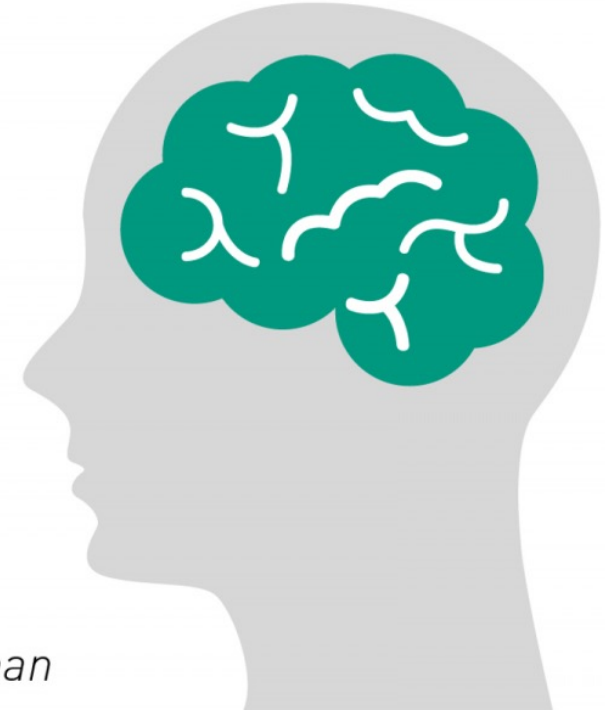
- A popular 2011 book by a Nobel prize-winning author who passed away last week ☹
- His model is that people have two different types of reasoning facilities
- **System 1** operates automatically and quickly, with little or no effort and no sense of voluntary control
- **System 2** allocates attention to effortful mental activities that demand it, including complex computations (e.g., logic, arithmetic, writing software, etc.)

SYSTEM 1

Intuition & instinct

95%

Unconscious
Fast
Associative
Automatic pilot



SYSTEM 2

Rational thinking

5%

Takes effort
Slow
Logical
Lazy
Indecisive

Source: Daniel Kahneman

Does that person look suspicious?

Who has the motive, means, and opportunity to do this?

Example question

Here is a simple puzzle

Don’t overthink it – give a quick answer

Example Question

Here is a simple puzzle

Don’t overthink it – give a quick answer

- **A bat and ball cost $1.10**
- **The bat costs one dollar more than the ball**
- **How much does the ball cost?**

Example Question

Here is a simple puzzle

Don’t overthink it – give a quick answer

- A bat and ball cost $1.10
- The bat costs one dollar more than the ball
- How much does the ball cost?

The ball costs 5 cents

- But many people will say 10 cents
- That’s our **system 1** at work!

Example Question

Here is a simple puzzle

Don’t overthink it – give a quick answer

- A bat and ball cost $1.10
- The bat costs one dollar more than the ball
- How much does the ball cost?

The ball costs 5 cents

- We get the correct answer we use our System 2 to use the math reasoning we learned in elementary school

Beyond today's chat systems

- Generative AI systems are part of the evolution of computer-assistive technology

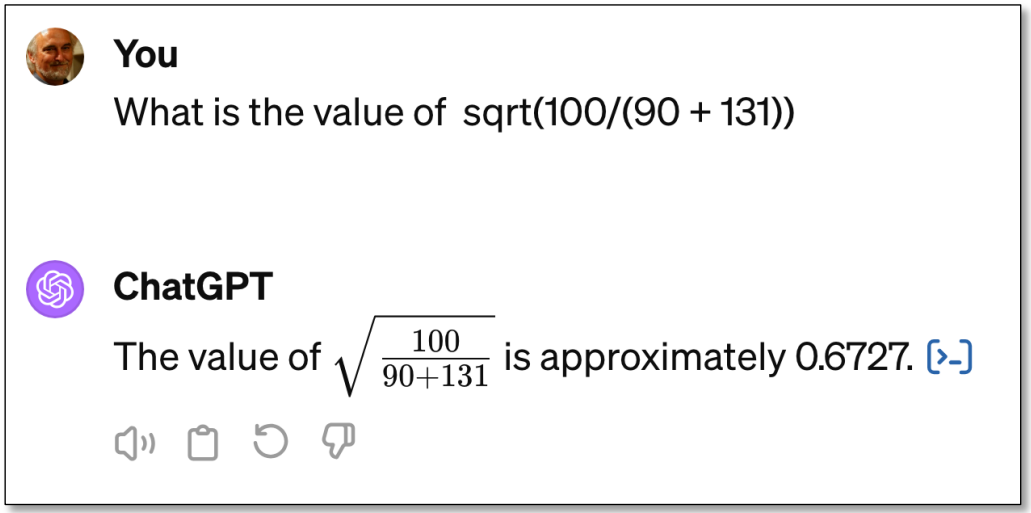
  Mechanical information retrieval ▸ computer IR ▸ web search ▸ web search with answers ▸ writing help (**Grammarly**) ▸ programming help (**GitHub Copilot**) ▸ …

- LLM size has **increased 10x** each year since 2018
- AI researchers working to identify & address shortcomings

  e.g., add links to provenance, common sense reasoning, structured knowledge, problem-solving, more logic and math, multilingual support, and more

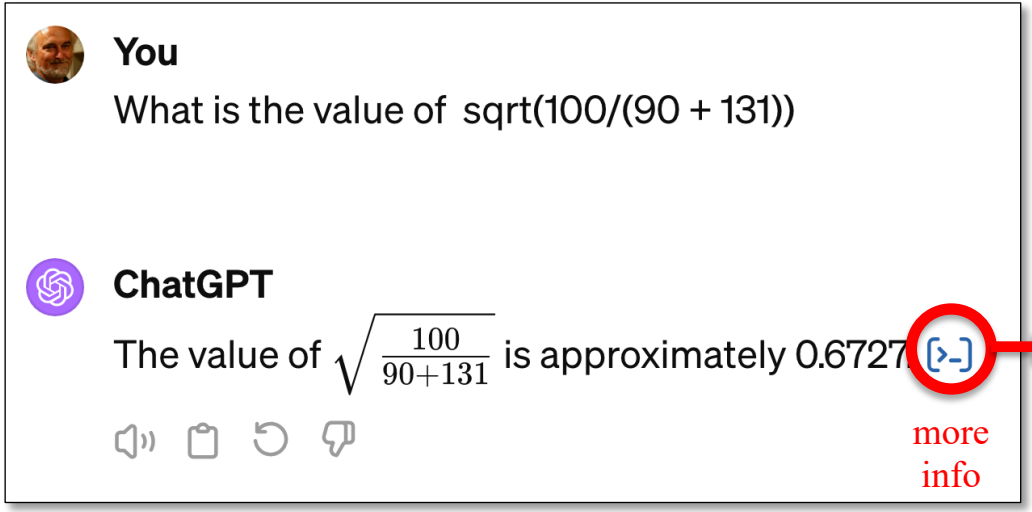
GAI systems can reason better

- Basic LLM-based systems are System 1 reasoners
- Current systems have added modules to handle many special System 2 tasks, e.g., math



GAI systems can reason better

- Basic LLM-based systems are System 1 reasoners
- Current systems have added modules to handle many special System 2 tasks, e.g., math



Retrieval-Augmented Generation (RAG)

RAG works for some queries by first retrieving information from an external knowledge source

1. User: What's the population of Belgium?
2. RAG component queries trusted external sources (e.g., Wikipedia) for relevant, current information
3. Retrieved text fed into model's generation component and used to create answer, e.g.:

   As of the latest statistics in 2021, the population of Belgium is approximately 11.6 million (source: Wikipedia)

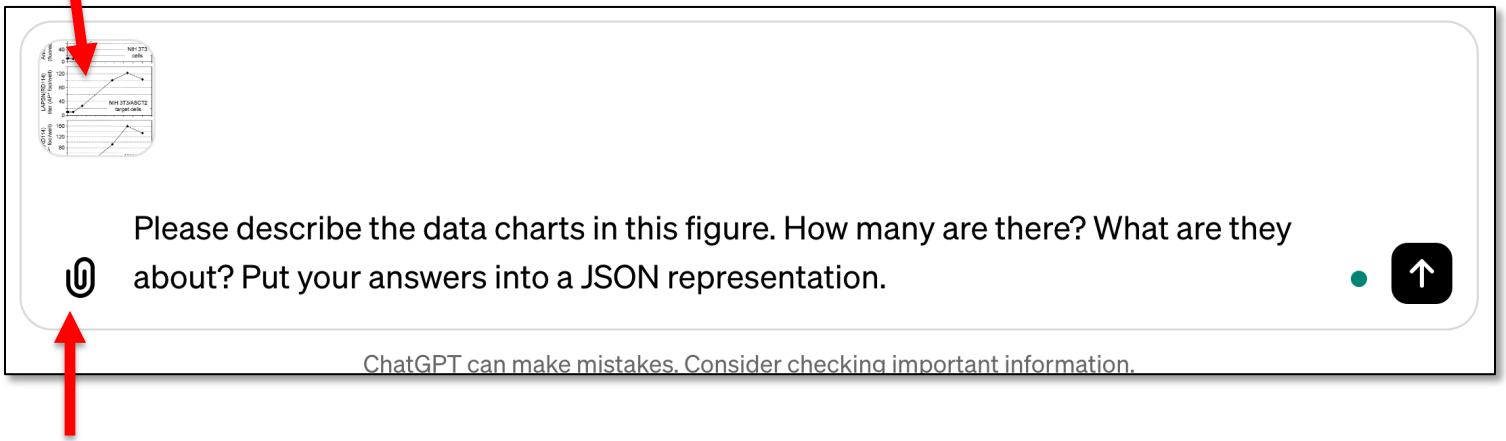
GAI systems can reason better

We can also get them to reason better by

- Prompt engineering: Give better instructions to define tasks, provide context, control style & tone, add constraints …
- Chain-of-thought prompting: guides the GAI through a step-by-step process, breaking it down into smaller, more manageable pieces

Extracting data from a figure

Image uploaded



I clicked to upload image from a file

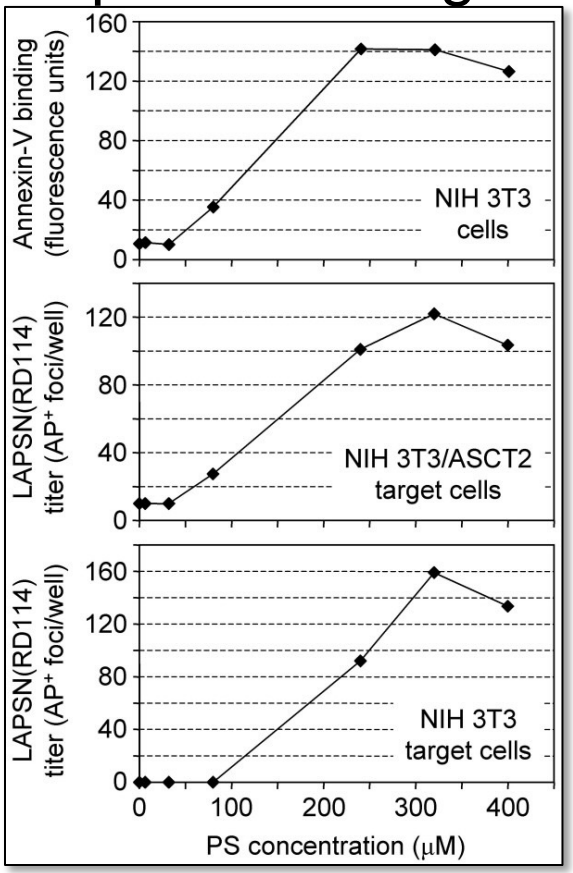
In the figure, there are three line charts, each showing the relationship between PS (phosphatidylserine) concentration in micromoles (μM) and some type of response measured in different units for NIH 3T3 cells or NIH 3T3/ASCT2 target cells. Here's the JSON representation of the charts:

json Copy code


```
{
  "charts": [
    {
      "title": "Annexin-V binding units",
      "x_axis": "PS concentration (μM)",
      "y_axis": "fluorescence units",
      "data_series": {
        "label": "NIH 3T3 cells",
        "data_points": [
          {"x": 0, "y": 0},
          {"x": 100, "y": 40},
          {"x": 200, "y": 80},
          {"x": 300, "y": 120},
          {"x": 400, "y": 160}
```


Each chart plots a series of data points that describe how the measured variable (either fluorescence units or titer) increases as the concentration of PS increases. The charts seem to depict a linear relationship, indicating that as PS concentration increases, the measured response also increases in a directly proportional manner. The actual data points are approximations based on the visual interpretation of the charts.

uploaded image



results

I asked ChatGPT to explain this cartoon

ChatGPT’s response

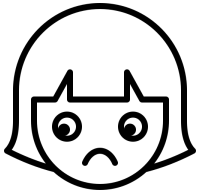
This cartoon is a humorous take on how artificial intelligence (AI) is used in the workplace to manage communication. On the left, an employee is bragging about how AI can expand a single bullet point into a long email that they can pretend to have written themselves. This suggests that the AI is doing the heavy lifting in terms of communication, allowing the person to seem more verbose or engaged than they actually are. On the right, another employee boasts about how AI condenses a long email into a single bullet point, which enables them to pretend they've read and comprehended the entire content.

Both panels are satirical comments on how AI can contribute to a lack of genuine communication in the office, with people relying on it to either inflate their own input or reduce their need to engage with others' messages.

More time would allow more examples

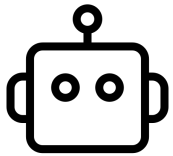
- Write, debug, or explain code
- Write an Excel macro
- Writing an SQL
- Extract info. from text &represent it as an RDF knowledge graph
- Image generation like this robotic True Grit based on a photo of the UMBC statue
- Summarizing a long document
- …

Some concerns



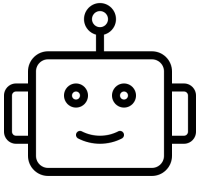
- GAIs used to create disinformation & deepfakes
- Negative impacts on society and climate
  - Eliminating jobs and disrupting education
  - Using too much energy to develop and deploy
  - Using private data w/o approval
- Misused for tasks that should be done by people
  - Take home exam!, writing and reviewing papers, …
- We might create an AGI that we could not control!

Artificial General Intelligence (AGI)



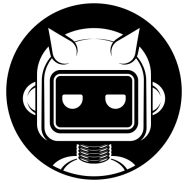
- An AGI is a system that can understand and do any intellectual task a human can
  - As opposed to narrow AI systems that do a few tasks
- AGIs are expected to have
  - Capacity to learn and adapt to new situations
  - Reasoning, problem-solving, & decision-making abilities
  - Potential to be conscious and self-aware
- AGIs are currently theoretical, but AI progress raises hopes and fears about them

AGI Potential Benefits



- Solving complex problems, e.g., help address major challenges in healthcare, climate change, poverty
- Automate and optimize tasks for economic growth and better quality of life
- Augment human intelligence and creativity, opening up new possibilities in science, art, and exploration
- Revolutionize medical diagnosis & treatment, leading to personalized medicine and increased lifespans

AGI Risks and Fears



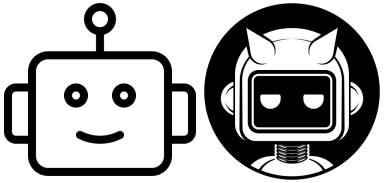
- AGI could have unintended and potentially harmful consequences that we cannot predict
- AGIs may become difficult or impossible to control, leading to existential risks
- AGIs will automate many jobs done by people, leading to unemployment & social unrest
- AGI might be used to develop autonomous weapons systems, threatening global security

Doomers, Boomers and P(doom)

- **AI doomers** have a very pessimistic view of AI's development and its implications for humanity
- **AI boomers** hold an optimistic view of AI and its potential benefits
- **P(doom):** probability estimate that unchecked AGIs cause existential catastrophes or human extinction, e.g.
  - Elon Musk: 20% - 30%
  - AI researcher Geoff Hinton: 10%
  - AI researcher Yann LeCun: <1%

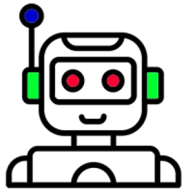
p(doom) examples from pause.ai site

Managing AGI



- AGIs may be inevitable
- Ethical considerations and responsible development are important, e.g.:
  - **Aligning AGI** development with human values and goals and adding appropriate **guardrails**
  - Ensuring AGI benefits are distributed fairly and equitably
- We need international collaboration and standards for open dialogue about potential risks and solutions

My perspective based on 50 years in AI

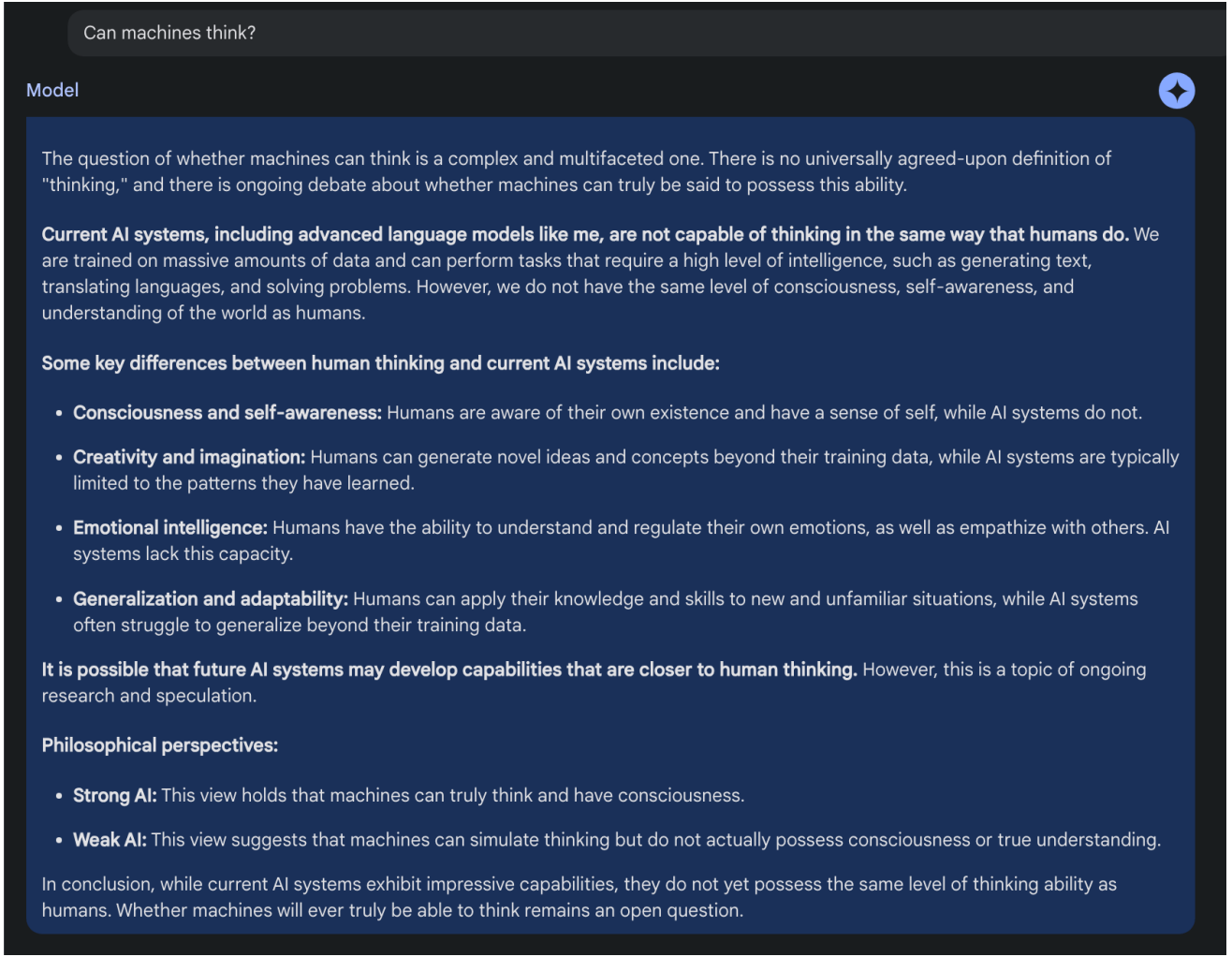


- We've not solved all of AI's problems nor yet developed what some call an AGI, **but good progress is being made**
- ChatGPT and similar systems (e.g., **Copilot**, **Gemini**, **Claude**) show remarkable and useful capabilities that
  - Are being integrated into software systems like web browsers, editors, programming environments, spreadsheets, and more
  - Can and will be improved by adding current and future AI advances
- Their **impact on society** will be like that of the **Web**, which was introduced about 30 years ago
- **Amara's Law**: "We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run"

Gemini's final word

Q: Can machines think?

A: …whether machines will ever truly be able to think remains an open question.



Thank you for listening

Questions?