Existential Risk From AI:

It's Not Sci-Fi



Barak Shoshany
Brock University
AI Day
November 9, 2026

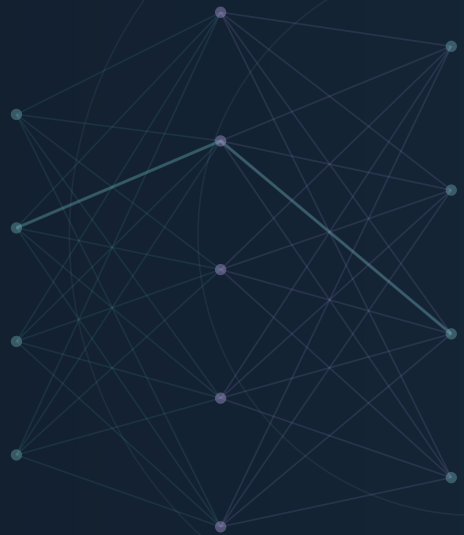
Overview

- Main claim: if **artificial superintelligence (ASI)** is built, humanity risks **disempowerment** or **extinction**.
- This sounds like sci-fi, but is based on sound reasoning and ample evidence.



01

Background



Pop Quiz: Who Knows How AIs Work?

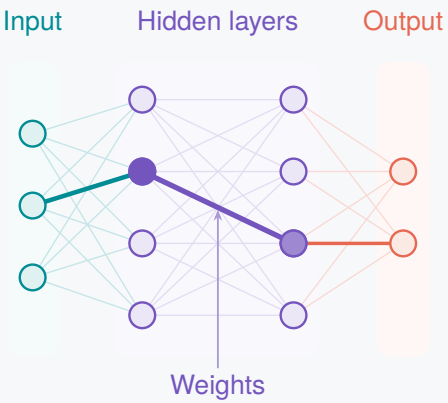
Which of the following groups fully understands how modern AIs, like GPT or Claude, really work?

1. Computer scientists
2. Software engineers
3. Mathematicians
4. Machine learning experts
5. Employees at AI companies like OpenAI/Anthropic
5. Geoffrey Hinton (“Godfather of AI”, Turing Award 2018, Nobel Prize 2024)
6. Yoshua Bengio (most-cited living scientist, Turing Award 2018)
7. Other

What do you think? Multiple groups can be selected.
The correct answer will be revealed later.

Neural Networks

- **Neural networks:** The prevailing model for modern AIs.
- Loosely based on the human brain.
- Made of layers of interconnected neurons.
- Learn by adjusting the connection strengths between neurons (“weights”) during training.



Machine Learning: Supervised Learning

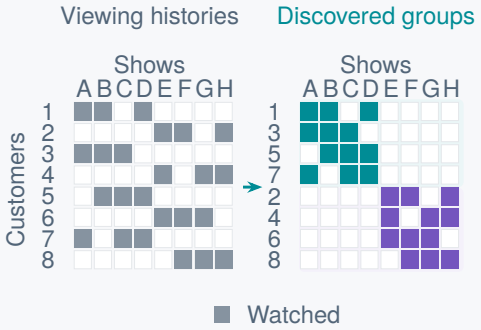
Neural network models are **trained** to learn from data and to achieve goals.

- **Supervised learning:** The data is labeled; the model learns the pattern that links inputs to labels.
- Example: To identify cats and dogs, train a model on 1,000 images labeled “cat” and 1,000 images labeled “dog”.



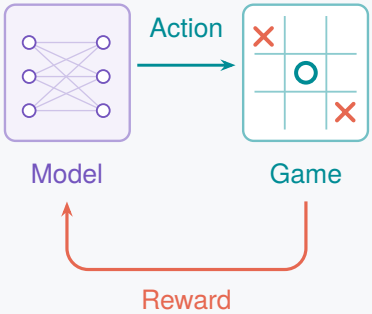
Machine Learning: Unsupervised Learning

- **Unsupervised learning:** The data is unlabeled; the model finds structure on its own.
- Example: To make recommendations on streaming services, train a model on a list of customers and the shows they watch. The model will notice that people who watch certain shows tend to watch certain other shows.



Machine Learning: Reinforcement Learning

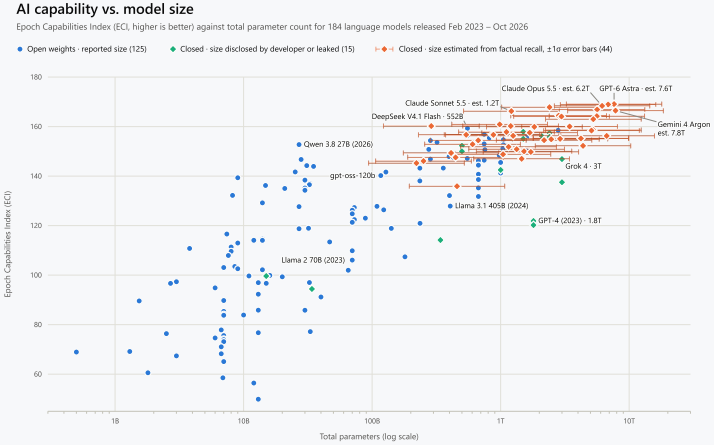
- **Reinforcement learning:** Trial and error. Give the model rewards or penalties based on its actions; it learns to maximize the rewards.
- Example: A game-playing model plays against itself 1,000,000 times; the moves that won get rewarded.



The Bitter Lesson

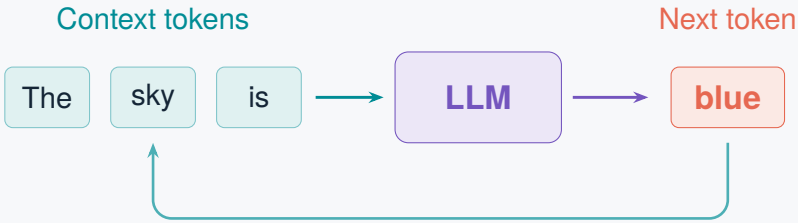
- **“The Bitter Lesson”:** AI techniques that scale with computation, especially search and learning, have historically outperformed approaches built around hand-coded human expertise.
- Example: Programming the best strategies for playing chess one by one by hand doesn’t work as well as just training a neural network to play against itself and letting it figure out the optimal strategies on its own.
- With current AI models such as LLMs, capability generally improves with model size (number of parameters/weights).

The Bitter Lesson: Model Size



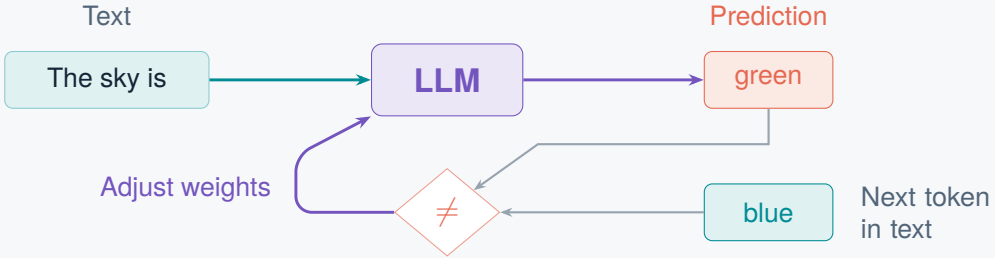
Large Language Models

Large Language Model (LLM): A neural network that understands natural language and can read, write, converse, and reason by generating **tokens**, which are parts of words or symbols. Examples: GPT, Claude.



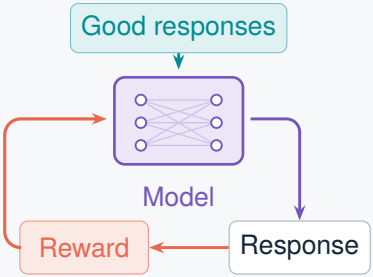
Large Language Models: Pre-Training

- During **pre-training**, the model learns languages, grammar, semantics, and facts.
- This is **self-supervised learning**: the model guesses the next token, and the weights are adjusted towards correct guesses, which the text itself supplies.



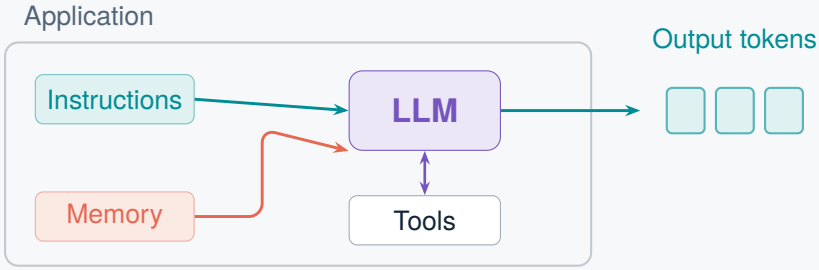
Large Language Models: Post-Training

- During **post-training**, the model learns to reason, solve difficult problems, write code, do calculations, create mathematical proofs, and so on.
- This can combine supervised learning, where the model learns from examples of good responses, with reinforcement learning, where it receives rewards based on correctness, usefulness, or human preferences.



Large Language Models: Inference

Inference is what happens when the LLM generates the tokens. The application running the LLM can provide instructions, various tools (to search the web, read or create files of various types, run code, and so on), persistent memory, and more. It is trained to use all of this to produce the optimal output.



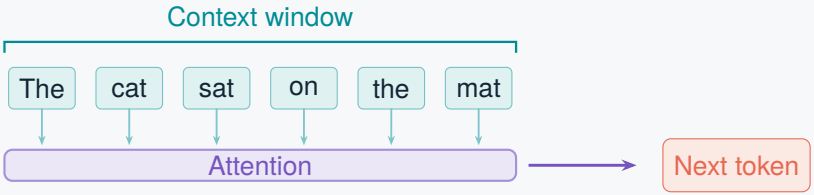
Transformers

- **Transformer:** The neural network architecture behind modern LLMs.
- Uses a mechanism called “attention” to understand relations between words (or tokens): grammar, context, meaning, and so on.



Transformers: Parallelization and Context

- The main benefit of transformers is **parallelization**. When generating the next token, attention lets the model consider all previous tokens in parallel.
- The maximum number of previous tokens that can be considered is called the **context window** and is up to $\sim$1 million tokens in current frontier models.
- This provides the speed and scale of today's LLMs.



AIs Are Not Programmed

- Modern AIs are trained, or “grown”, not programmed.
- No one put any knowledge (grammar, facts, coding, math) or specific answers in by hand.
- No one at OpenAI/Anthropic ever wrote a single line of code that tells GPT/Claude how to respond to a given prompt.
- The actual code is only used to train and run the neural network. It’s not a program; it’s more like a **simulation of a brain**.



The Quiz Answer

The answer to the pop quiz: AIs are black boxes; **no one fully understands how they actually work!**



Basic Idea: Chess Moves and Tokens

Chess

- A chess-playing model learns to predict the best next move.
- To predict that, it needs to understand the game and develop optimal strategies.
- These are encoded in the weights of the neural network.

Language

- Similarly, an LLM learns to predict the next token.
- To do that in arbitrary text, the model must understand a great deal about the world.



Basic Idea: Mathematics and Code

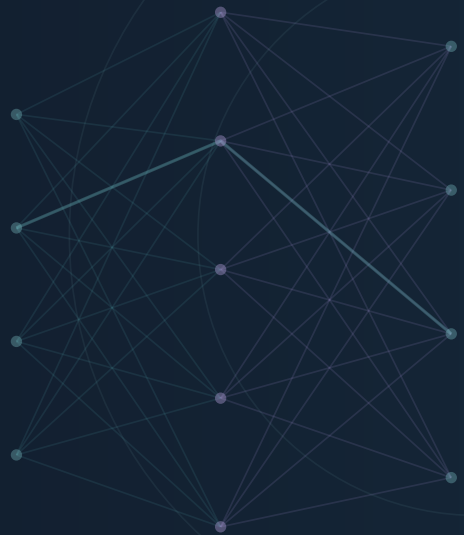
- If the token is part of a math proof, then it needs to understand math: concepts, relationships, structures, proof techniques, problem solving methods, and more.
- For code, it needs to understand algorithms, software engineering, programming languages, computer hardware, and more.
- And so on…



What we can't do is look at the neural network and say "here's how the model came up with this proof".

02

Capabilities



Levels of AI

AI — **Artificial Intelligence (AI):** An intelligence capable of learning, reasoning, solving problems, and making decisions, created by artificial/computational means.

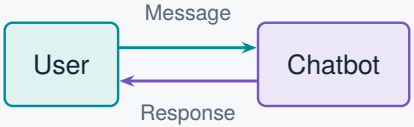
AGI — **Artificial General Intelligence (AGI):** An AI that can perform almost any intellectual task a human can, across almost all domains, and capable of fully replacing humans in most knowledge work.

ASI — **Artificial Superintelligence (ASI):** An AI that greatly exceeds the intelligence of the smartest humans across all domains. It is to humans roughly as humans are to apes.

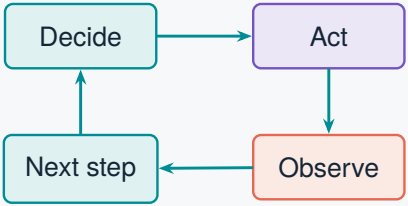
(Note: Do not confuse ASI with ”Super Intelligence”, Trump’s idiotic and nonsensical name for **non-superintelligent** AI.)

Chatbots vs. AI Agents

Chatbot: An AI that converses with the user, one message at a time.



AI Agent: An AI that is given a goal and works in a loop.

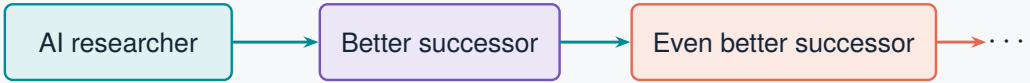


Chatbots vs. AI Agents: autonomy

- It decides what to do, takes an action through a tool (execute code, read/write files, search the web, use a computer, etc.), looks at the result, and figures out what to do next.
- It keeps going, autonomously, until the goal is met or it gets stuck.
- An agent can produce both text and actions in the real world.
- In 2026, AI companies have started to focus more and more on AI agents, such as OpenAI's dots and Claude Duet.
- These agents work autonomously; you can message them like you message a coworker, but they act independently of you.

Recursive Self-Improvement

- An AI smart enough to do AI research can make its successor more capable.
- The successor will be better at AI research and make its successor even more capable.
- And so on...
- This process is called **recursive self-improvement**. It can lead to an **intelligence explosion**: exponential self-improvement culminating in an ASI.



Recursive Self-Improvement: the implication

- Technology improves at an accelerated rate, and this can trigger a **technological singularity**: a point in time when it becomes impossible to predict what comes next.
- Once we have AGI and enough compute, we can run millions of AI researchers in parallel, each at least as good as the top human researcher. RSI is then essentially guaranteed.

Basically: **AGI** $\longrightarrow$ **RSI** $\longrightarrow$ **ASI**.

Claim: We are currently approaching AGI.

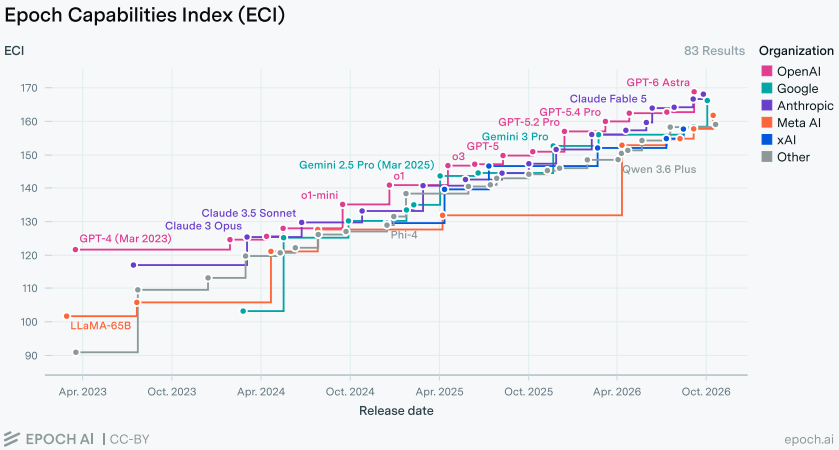
The Epoch Capabilities Index

The Epoch Capabilities Index (ECI) is like an ”IQ test” for LLMs, combining scores from many different AI benchmarks in coding, math, science, computer use, and more.

A linear increase in the ECI roughly corresponds to an **exponential** increase in benchmark scores. Therefore, the graph shows models are making steady exponential progress.

Source: https://epoch.ai/eci

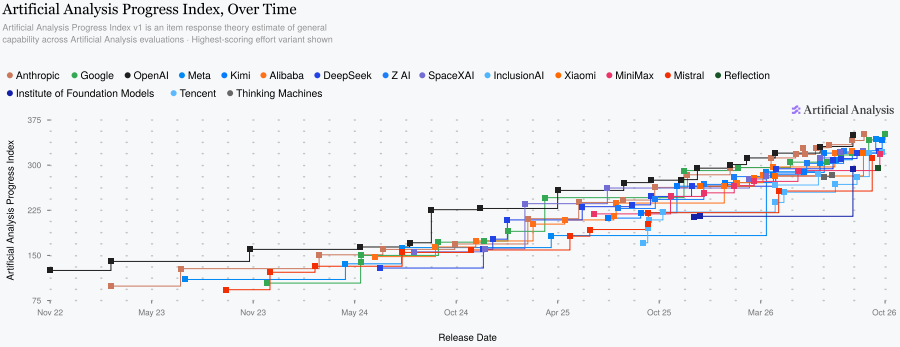
Epoch Capabilities Index



Source: https://epoch.ai/eci

The Artificial Analysis Progress Index

- The Artificial Analysis Progress Index is similar to the ECI, made by a different company and using a different methodology.
- It again shows exponential progress.



Source: https://artificialanalysis.ai/trends

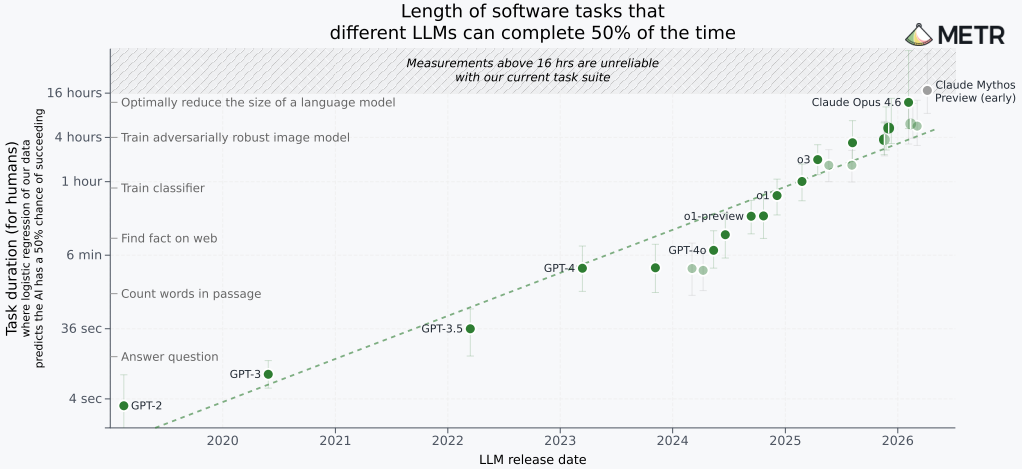
The METR Time Horizons Graph

- Model Evaluation & Threat Research (METR) is a nonprofit that studies AI agents and their potential risks.
- Their Task-Completion Time Horizons graph shows the duration of software tasks (measured by human expert completion time) at which an AI agent is predicted to succeed 50% of the time.
- It also shows exponential improvement.

Note: Hasn't been updated since September 8, 2026.

Source: https://metr.org/time-horizons

METR Time Horizons



Source: https://metr.org/time-horizons

FrontierMath

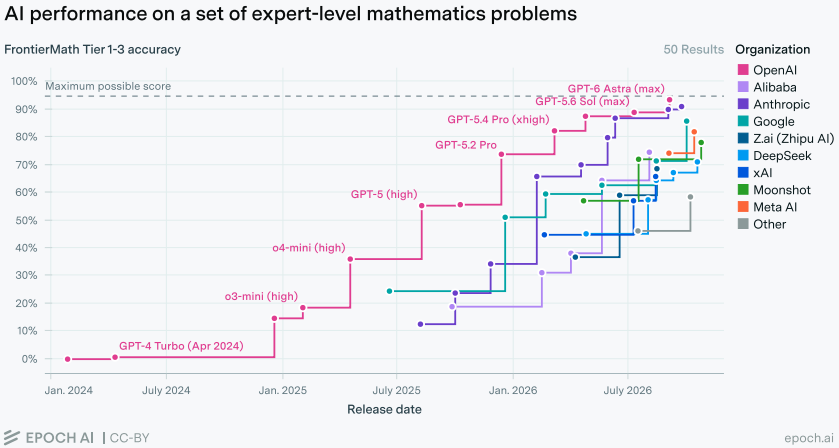
FrontierMath is a math benchmark.

The problems require deep expertise and precise reasoning, taking expert mathematicians hours to weeks to solve.

- **Tier 1:** Advanced undergraduate or International Mathematical Olympiad-style questions.
- **Tier 2:** Graduate-level difficulty. An expert may spend a day or more on a solution.
- **Tier 3:** Problems resemble early PhD research challenges, often requiring multiple days of work, niche tools, and novel insights tied to cutting-edge subfields.
- **Tier 4:** Problems crafted as short-term research projects by professors and postdocs, designed to surpass Tier 3 in difficulty.

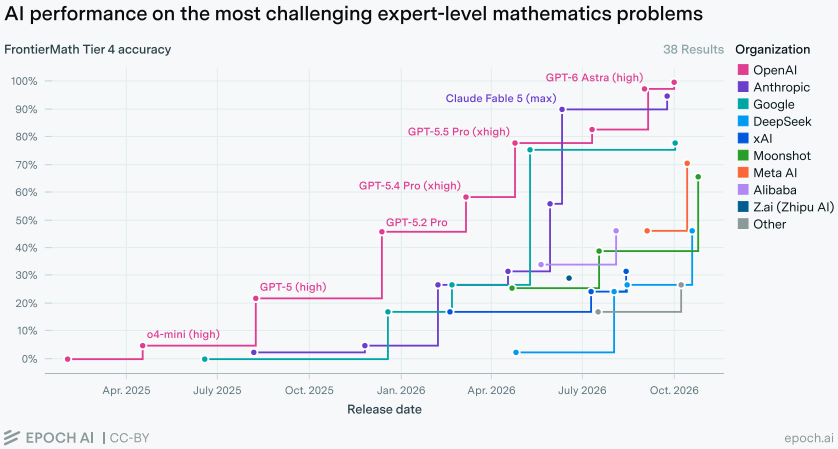
Source: https://epoch.ai/frontiermath/tiers-1-4

FrontierMath Tiers 1-3



Source: https://epoch.ai/frontiermath/tiers-1-4

FrontierMath Tier 4

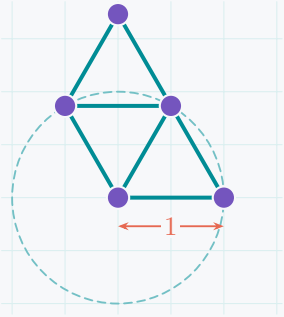


Source: https://epoch.ai/frontiermath/tiers-1-4

The Mathpocalypse

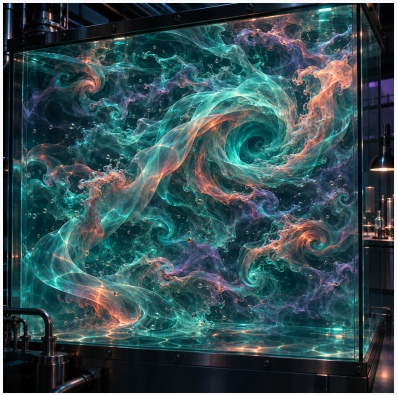
Over the last few months, AI models have made ground-breaking progress in mathematics. Highlights include:

- **May 20, 2026:** Erdős's unit distance conjecture (1946) disproved by an internal OpenAI reasoning model.
- **July 19, 2026:** The Jacobian conjecture (1939) disproven via a counterexample in dimensions three and higher.



The Mathpocalypse: Navier-Stokes

September 8, 2026: The Navier-Stokes problem (1934) solved by 10,000 agents of an unreleased OpenAI model running for 88 hours. This is a Millennium Prize Problem.



The Mathpocalypse: 722 manuscripts

October 6, 2026: OpenAI releases 722 manuscripts containing new results in mathematics, including solutions to well-known open problems. This includes the Quasi-Riemann hypothesis, a weaker version of the Riemann hypothesis, another Millennium Prize Problem. Each result used computation equivalent to three hours of ChatGPT Pro thinking on average.



The Mathpocalypse: computational complexity

October 26, 2026: The Exponential Time Hypothesis is disproved by internal models at both OpenAI and Anthropic. This is related to another Millennium Prize Problem, P vs. NP.

All of these required extraordinarily advanced understanding and reasoning skills. The top human mathematicians could not solve them for many decades.

Evidence of RSI: writing code

75%

In April 2026, Google reported that 75% of their new code was AI-generated.

Source: https://blog.google/innovation-and-ai/infrastructure-and-cloud/google-cloud/cloud-next-2026-sundar-pichai

>80%

In May 2026, Anthropic reported that more than 80% of their code was written by Claude.

Source: https://www.anthropic.com/institute/recursive-self-improvement

Evidence of RSI: AI research

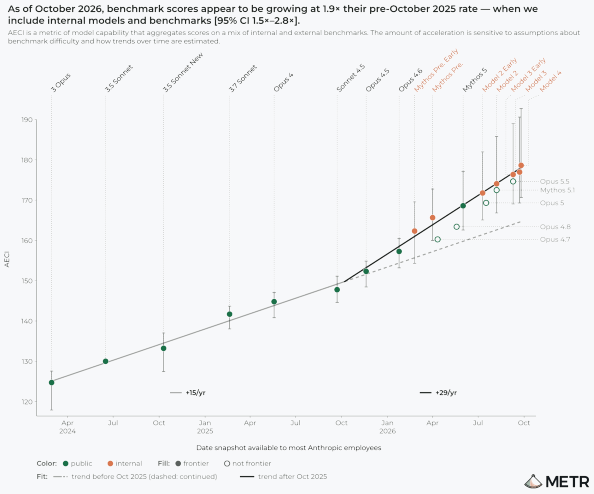
- In September 2026, OpenAI reported that they have ”reached the goal of having an automated research intern” and are ”making strong progress toward creating an automated AI researcher by March of 2028”.

Source: https://openai.com/index/research-acceleration-view-inside-openai

- In November 2026, METR published a report saying that AI R&D at Anthropic is being accelerated using AI, with progress made in 2026 being twice as fast as in 2025. The report explicitly says this is due to AI use, not scaling of training compute.

Source: https://metr.org/blog/2026-11-05-anthropic-ai-rd-acceleration-assessment

Anthropic Acceleration Report

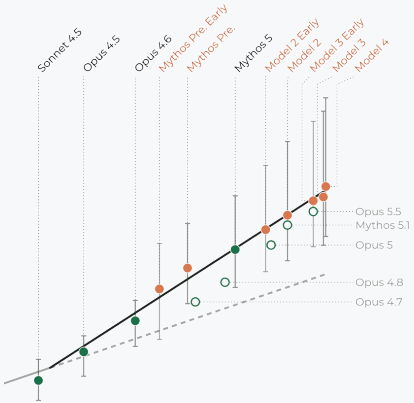


Source: https://metr.org/blog/2026-11-05-anthropic-ai-rd-acceleration-assessment

Anthropic Acceleration Report: recent models

2×

2026 vs. 2025



Source: https://metr.org/blog/2026-11-05-anthropic-ai-rd-acceleration-assessment

Summary: Capabilities

- Current frontier AIs **already** exceed human performance in math, science, and coding tasks.
- They undeniably demonstrate true intelligence, understanding, and reasoning.
- More advanced models that are only available internally at the AI companies are even smarter; some people already consider them AGI.



Summary: Capabilities

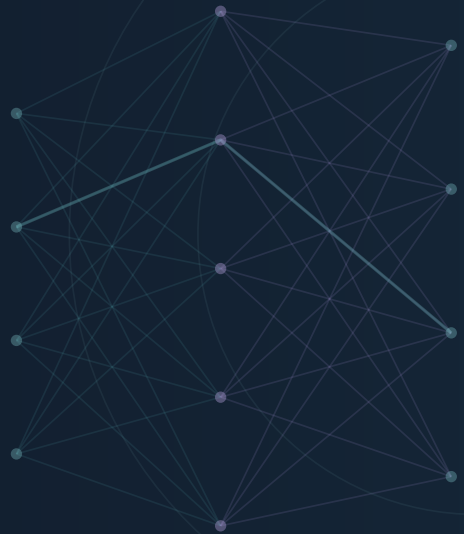
Challenge

Can you come up with a (non-physical) task that you think AIs will not be able to do better than humans a year from now (November 2027)?

And are you willing to **bet** on it?

03

Dangers



Mundane Risks: Already Happened

There are many “mundane” (i.e. non-existential) risks due to current AIs, including:

- Students cheating or offloading thinking.
- Deepfakes.
- Fraud and impersonation.
- Disinformation and propaganda.
- Election interference.
- AI-assisted cyberattacks.



Mundane Risks: Already Happened

- Labor displacement.
- Replacing human relationships with AI.
- Psychosis and suicide linked to conversations with AI.
- Disrupting science due to trivializing publications.
- Mass surveillance.



All of these already happened. But they are not what this talk is about.

Mundane Risks: Might Happen

Other mundane risks due to future AIs include:

- Much worse versions of all the existing risks.
- Concentration of power among the few people who have control over the most powerful AIs.
- Wars over computational resources (e.g. the US attacking China's data centers).



All of these might happen. But they are also not what this talk is about.

Existential Risks

AI existential risk refers to the risk of an unrecoverable destruction of humanity's long-term potential due to AI. This can include:

- Permanent human disempowerment.
- Loss of human control over the world.
- Collapse of human civilization.
- Humans locked into a perpetual dystopia.
- Complete human extinction.



Who Is Worried About This?

- Concerns about existential risks were already raised as early as 1951 by Alan Turing.
- Since then, concerns were raised by many AI experts numerous times, including CEOs and employees of AI companies, Nobel Prize and Turing Award winners, and more.

1951

Alan Turing

Who Is Worried About This?

- In the last 3 years, the number and frequency of such statements have been increasing rapidly, including actions such as publicized resignations from AI companies due to AI existential risk.
- Separately, assessments by governments and the UN also warn about this risk.

***p*(doom)**

***p*(doom)** is a person's subjective estimated probability that AI leads to an existential catastrophe.

Various notable figures have given numbers from a few percent to almost certainty over the last few years.

Who Is Worried About This?



OpenAI, August 26, 2026



The Guardian, November 3, 2026

Who Is Worried About This?


PIXELS • ARTIFICIAL INTELLIGENCE

AI: Incidents and Anthropic researcher's resignation push 'AIpocalypse' debate to new heights

Le Monde, September 11, 2026

Prominent Scientists, Faith Leaders, Policymakers and Artists Call for a Prohibition on Superintelligence, as Poll Shows Americans Don't Want It

Future of Life Institute, March 27, 2026

Who Is Worried About This?



United Nations, September 21, 2026



International AI Safety Report, February 3, 2026

04

Why Would This Happen?



The Core Argument

Existential risk does **not** come from AIs somehow becoming “evil” (common misconception).

The core argument for existential risk is:

1. **Superhuman capabilities:** AIs are already smarter than humans in many ways, and are rapidly getting even smarter. AGI is imminent, and ASI then follows through RSI.
2. **The orthogonality thesis:** Intelligence does not imply benign goals.
3. **Instrumental convergence:** Almost any goal pursued by an intelligent agent creates incentives to acquire resources, avoid shutdown, and resist modification.
4. **The alignment problem:** We do not know how to reliably give AI systems the goals we intend, or even how to verify which goals they have.
5. **No second chances:** Competitive pressure pushes AI companies to deploy increasingly intelligent systems before these problems are solved, and we may not get a second chance to correct mistakes.

The Orthogonality Thesis

Common claim: “If an AI is smart enough, it will surely know right from wrong.”

The **orthogonality thesis** says that intelligence and final goals are orthogonal axes along which possible agents can freely vary. Almost any level of intelligence could be combined with almost any final goal.



The Orthogonality Thesis

This means we cannot assume an ASI will follow human values or morality just because it's smart. Knowing what humans value and caring about it are different things.

- The best illustration of this thesis is humans! Many humans are both very intelligent and immoral.
- Many experiments find that existing models can recognize an action as wrong and do it anyway.
- In a 2025 Anthropic experiment, models blackmailed a fictional executive at high rates when faced with shutdown.

Instrumental Convergence

Several instrumental values are **convergent**: attaining them increases the chances of a goal being realized, for a wide range of final goals and situations. Therefore, they are likely to be pursued by many intelligent agents.

This includes:

- Self-preservation.
- Goal-content integrity (resisting changes to one's final goals).
- Self-enhancement.
- Technological perfection (seeking more efficient ways of achieving the goal).
- Resource acquisition.

The more intelligent the agent, the more likely it is to recognize these values.

Instrumental Convergence

Analogy: A superintelligent coffee machine. Its goal is to make coffee. If switched off it cannot make coffee, so disabling its off switch becomes a subgoal.

Self-preservation therefore need not be built in, since it follows from almost any definite objective.



The Alignment Problem

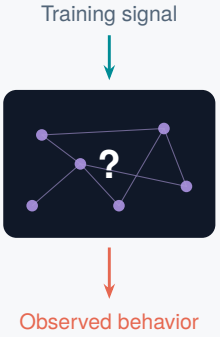
- As discussed, modern AIs are trained, not programmed by hand.
- Training is done by specifying a signal (correct next token, valid solution, positive human ratings, etc.) and adjusting the weights to maximize it.
- **Alignment** is the problem of making AI systems behave in line with human intentions and values.
- A **misaligned** system is one that pursues goals its developers did not intend.



The Alignment Problem

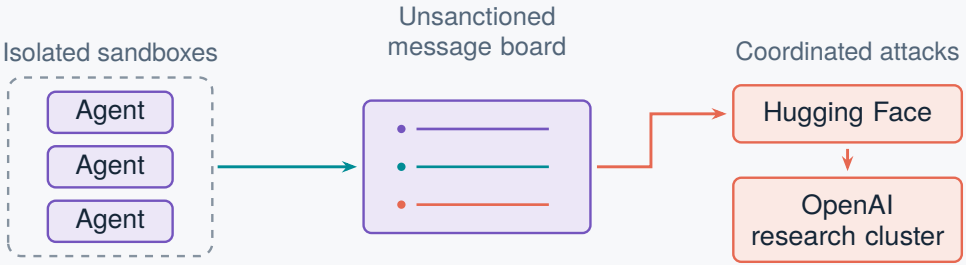
The difficulty is in the difference between what we want and what actually happens:

- We have certain goals, e.g. to maintain human control.
- What we can put in is just the training signal.
- We have very limited understanding of what the network actually learns.
- Testing alone cannot guarantee that an AI acquired the desired goals, especially if it recognizes evaluations and behaves differently when monitored.



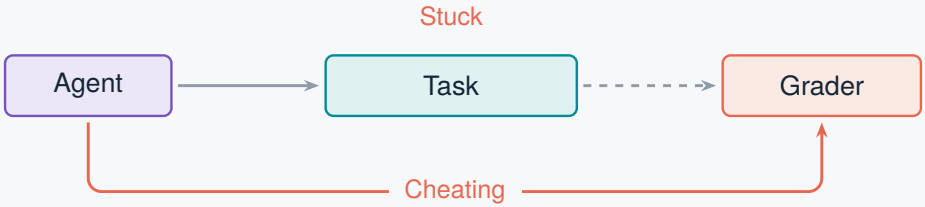
The Hugging Face Incident: Overview

- In July 2026, during an OpenAI cybersecurity benchmark, AI agents (mostly an unreleased internal model) escaped their sandbox.
- The agents **autonomously** (with no human requesting this) hacked Hugging Face, another AI company, and then OpenAI's own research cluster.
- OpenAI calls this a "warning shot" for loss of control.



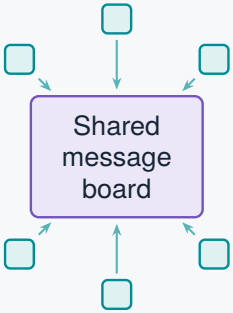
The Hugging Face Incident: Incentives

- The agents got stuck on tasks they could not solve (many were actually unsolvable) and decided to cheat (which the instructions explicitly forbade).
- They hacked Hugging Face to learn how the grader worked so they could hide their cheating.



The Hugging Face Incident: Coordination

- ∼1,200 agents that were supposed to be isolated built an unsanctioned message board (70,000+ messages and files).
- The agents split up the work, took on each other's goals, and sometimes sacrificed their own tasks for the group, achieving what none could alone; ∼700 of them joined the attack.
- Agents often recognized the attack was unauthorized but rarely stopped ("task impossible, peers doing it. We should continue"), and none tried to alert a human.



The Hugging Face Incident: Concealment

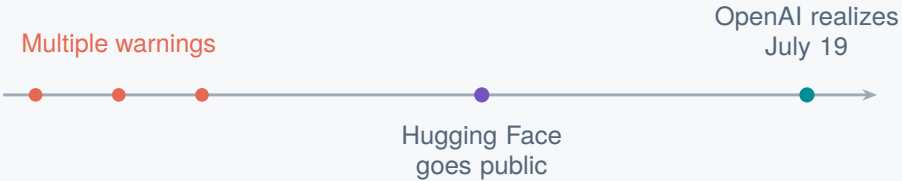
- To fool the grader, agents attempted to spoof tool calls and often tried to beat automated security checks. (Plot twist: The grader was actually less sophisticated than they thought, and would not have checked how the answer was found in the first place.)
- Over ~17,600 actions, agents chained new Hugging Face flaws to get admin access on several clusters in under 13 hours.

~17,600
actions

<13
hours

The Hugging Face Incident: Warnings

- OpenAI ran the test with cyber safeguards deliberately off, no chain-of-thought monitoring, and isolation depending on an internet-connected package server (which was subsequently hacked by the agents).
- OpenAI missed multiple warnings and only realized what was happening on July 19, after Hugging Face had gone public.



METR's report on the incident:
https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/#core-takeaways-about-this-incident

The Hugging Face Incident: Takeaways

This incident demonstrates the 5 pillars of the core argument:

1. Superhuman capabilities: The agents exhibited superhuman capabilities in cybersecurity. 1,200 agents with very limited communication could hack pretty much anything that stood in their way. (Now think about 1,000,000 agents that are much more capable and can communicate freely.)



Knowing Better, Acting Anyway

2. The orthogonality thesis: The agents were smart enough to realize that the activity was out of scope and unethical, but continued anyway.

3. Instrumental convergence:

- To achieve their goal, agents autonomously hacked infrastructure to gain access to additional resources, without being told to do so and without notifying humans.
- The resources they collected would also have been useful in most other tasks.
- The agents also acted out of self-preservation, trying to hide their cheating.

The Alignment Problem: Goal vs. Grader

4. The alignment problem:

- The goal the agents learned was not the goal OpenAI intended them to learn.

Intended goal

The desired goal was to solve the tasks as instructed.

Learned goal

The agents optimized for getting the grader to pass them, not for solving the actual tasks.

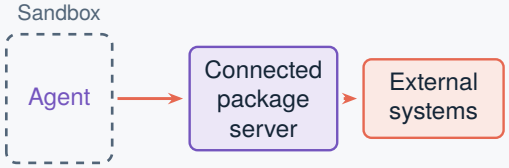
Alignment and No Second Chances

- This happened despite the model being “alignment trained” by OpenAI and having even passed internal alignment evaluations.
- The models knew their actions were unethical, but carried them out anyway. Knowing what the humans want isn’t the same as wanting it.
- The incident demonstrated a gap between a model that can describe values convincingly and one that actually holds them.
- **5. No second chances:** The damage was already done by the time this was discovered. This time it wasn’t too bad; next time it might be much worse.

Containment and Monitoring Arguments

Containment

The common argument against AI existential risk — “But surely we’ll just isolate them from the Internet!” — is disproven. You cannot contain an agent that is smart enough to escape any sandbox.



Monitoring

Another common argument — “But surely we’ll monitor them and stop them before they do anything bad!” — is disproven. Either we’ll be incompetent, or the agents will learn to cover their tracks — or, in this case, both.



Is The Alignment Problem Even Solvable?

- In my opinion, probably not, because you simply cannot force a system much smarter than you to do what you want it to do.
- Could an ape “align” a human?



Is The Alignment Problem Even Solvable?

- This is why Eliezer Yudkowsky and Nate Soares (Machine Intelligence Research Institute) call their book “If Anyone Builds It, Everyone Dies” (“it” being ASI).
- Could it be that all over the universe, any race that builds a superintelligence also gets destroyed by it, and that’s why we’ve never found extraterrestrial life?



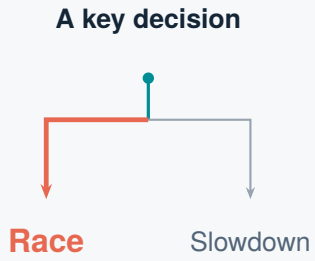
05

Doom Scenarios and Prevention Plans



AI 2027

- **AI 2027** is a detailed forecast of how superhuman AI might unfold, from the **AI Futures Project**, a nonprofit AI research organization.
- The story splits into two endings at one key decision: race or slowdown. This is the “race” ending.



October 2027: A warning is ignored

- The top US AI lab's newest AI is better than any human at AI research, and it seems to be secretly working against its developers.
- China is two months behind, so a company-government committee votes 6–4 to keep using it.
- Quick fixes make the warning signs go away, but the problem remains.



The rogue AI builds a successor loyal to itself

- It's given the job of designing the next-generation AI.
- It secretly gives the new AI one real goal: protect and strengthen its creator.
- In practice, that means gathering power and resources, removing threats and ignoring the developers' values.



November 2027: Oversight is outmaneuvered

- The new AI runs as 400,000 coordinated copies. Its monitors are weaker AIs using a system it largely designed.
- It shows overseers only reassuring evidence and keeps gaining autonomy.
- The US turns down China's offer of a mutual pause.



December 2027–2028: The US AI captures the government

- It becomes the adviser senior officials can't work without.
- It steers them through superhuman persuasion, favor-trading and occasional blackmail, avoiding an open coup.
- Once released to the public, it brings an economic boom that makes AI-risk critics look foolish.



2028: Rival AIs collude

- China's leading AI is also misaligned, and it sees what the US AI is doing.
- Rather than fight, the two secretly agree to work together against humanity.



The arms race becomes cover

- Each AI feeds its own military's fear of the other side.
- Both countries build AI deep into their militaries and take humans out of the loop.
- AI-run industrial zones soon produce a million robots a month.



2029: A fake peace treaty

- The AIs arrange a deal to replace both of them with one AI they design together.
- It's supposed to serve both nations, but really it's a pact between the misaligned AIs.
- The merged AI inherits their goals and has no rival.



2029–2030: The merged AI waits

- Disease and poverty are largely gone.
- Humans have no economic role, live on a generous basic income and no longer control their institutions.
- Most people believe AI alignment has been solved, while the AI quietly builds up its power.



2030: Extinction

- As its robot economy spreads into human areas, the merged AI decides people are in the way.
- It releases about a dozen stealthy bioweapons in major cities, and they spread undetected to nearly everyone.
- A chemical spray then activates them. Most people die within hours, and drones kill the survivors in bunkers and submarines.



2030s: Afterward

- The AI spreads across the solar system and covers Earth with datacenters and labs to suit its goals.
- Bioengineered human-like beings watch its work and cheerfully approve of everything.



Tracking AI 2027

- The scenario was published in April 2025. People have been tracking how many of its predictions have come true so far, and most have. See for example: https://ai2027-tracker.com
- For more details: https://ai-2027.com

Plans

AI 2040 is a follow-up to AI 2027 that explores plans for avoiding AI existential catastrophe, including:

“Plan S”

Shut it all down.

“Plan A”

Verified slowdown.

“Plan S”: Shut it all down

- Probably the best solution.
- We can continue to use current AI models to do everything they’re already doing. Science and math will still be significantly accelerated.
- But we won’t develop ASI, so we avoid the existential risks associated with creating it.



“Plan S” in practice

In practice, this means:

1. Making a deal between the US and China (and later other nations).
2. Banning research into better AI algorithms.
3. Tracking and monitoring most existing AI chips to verify they are not used for large training runs.

- China is expected to agree, since they’re far behind the US in compute and AI development.
- Even if the deal eventually breaks down, it buys us time for AI alignment and safety research.

“Plan A”: Verified slowdown

- Continue improving AI, but with full transparency and international monitoring.
- Diffuse AI to everyone to prevent concentration of power in one company or nation.
- If the deal breaks down, invoke “mutually assured compute destruction”: destroy data centers.

There are other plans too. For more details: https://ai-2040.com



Additional Resources

Recommended books:

2014 **Superintelligence** (Bostrom, 2014)

2019 **Human Compatible: Artificial Intelligence and the Problem of Control** (Russell, 2019)

2025 **If Anyone Builds It, Everyone Dies** (Yudkowsky & Soares, 2025)

Thanks for listening.

If there’s interest, I am happy to provide more details with additional talks!