Remarks on AI Development Timelines and Risks

Joshua Achiam

10.21.26

About Me

- B.S. Aerospace Engineering, B.S. Physics
  University of Florida, 2014

- PhD in Electrical Engineering and Computer Science
  UC Berkeley, 2021
  - Thesis: “Exploration and Safety in Deep Reinforcement Learning”

- OpenAI July 2017 - July 2026
  - AI safety research scientist, primarily early reinforcement learning research
  - Head of Mission Alignment, 2024-2026
  - Chief Futurist, 2026

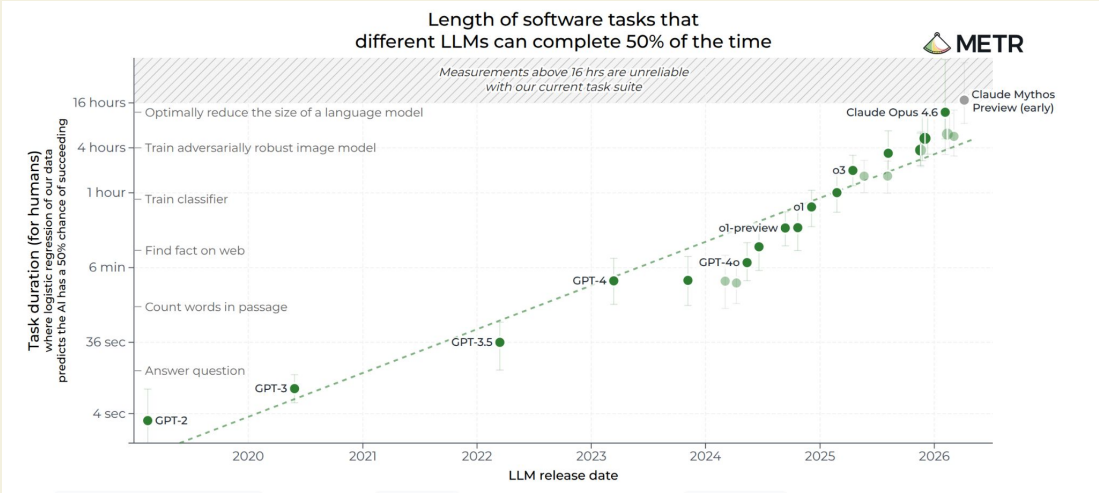
- Independent! July 2026 - Present

- Starting Oct 2026:
  - Non-Resident Senior Fellow, Institute for Progress
  - Non-Resident Senior Fellow, Foundation for American Innovation



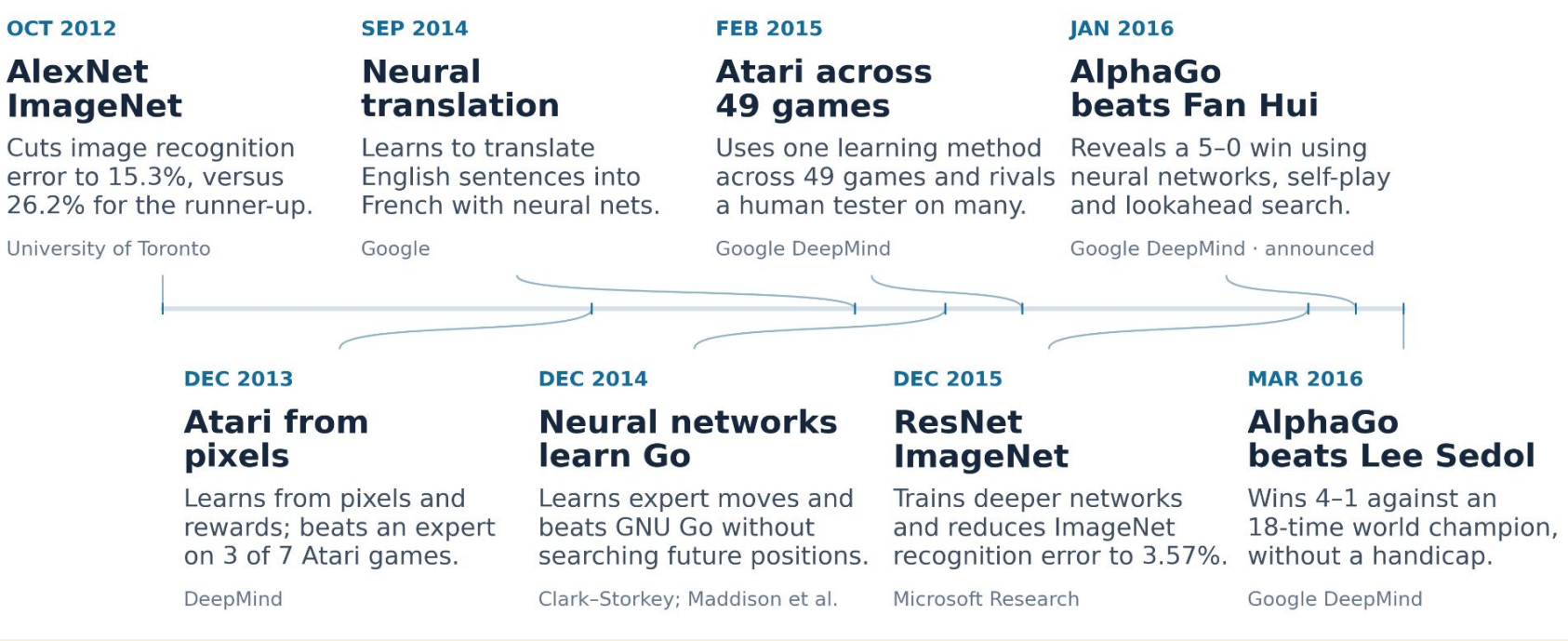
Superintelligence and Recursive Self-Improvement

- It's easy to spend too much time on definitions; I'll try not to
- A handwave: AI is superintelligent in a domain if it's better than the best humans at something
- The frontier AI models we have today are probably already superintelligent in some interesting ways
- We should anticipate that model capabilities will get stronger; the consensus in the field is that things are accelerating.
- The field considers RSI a near-term (next 2 years) phenomenon.

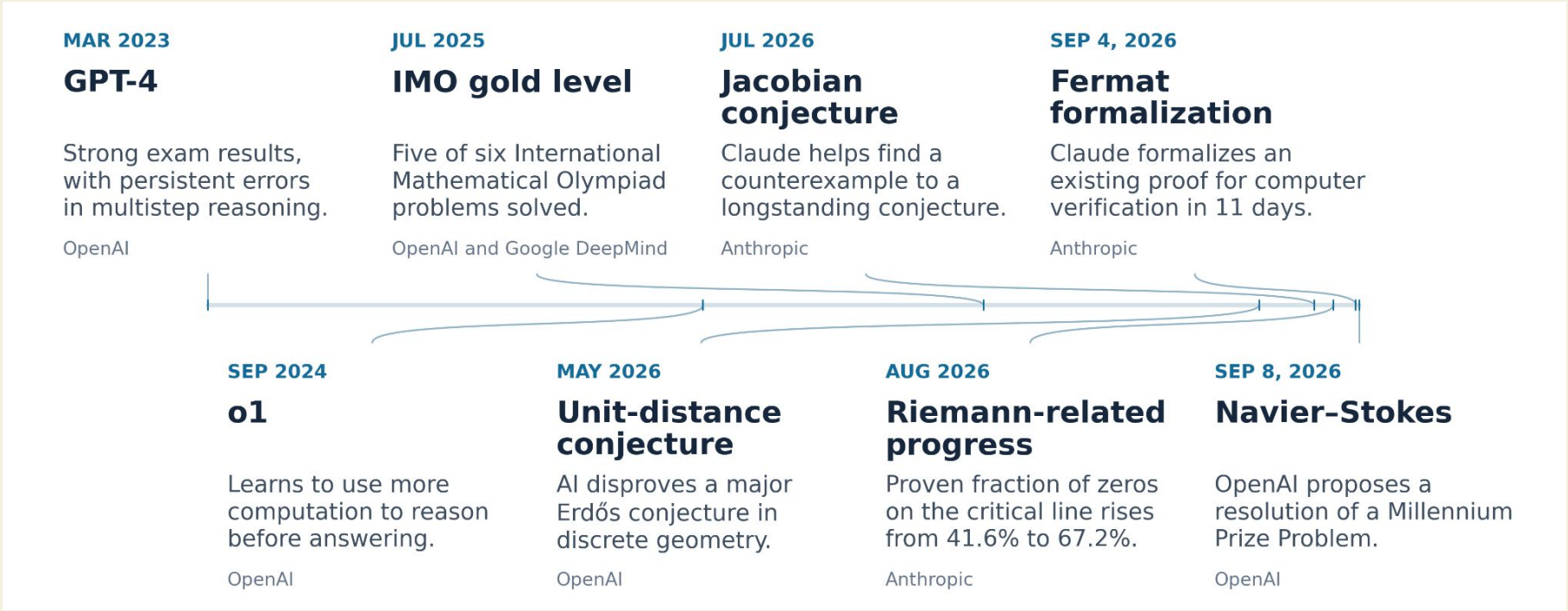


https://metr.org/time-horizons/

AlphaGo


OCT 2012
AlexNet ImageNet
Cuts image recognition error to 15.3%, versus 26.2% for the runner-up.
University of Toronto
SEP 2014
Neural translation
Learns to translate English sentences into French with neural nets.
Google
FEB 2015
Atari across 49 games
Uses one learning method across 49 games and rivals a human tester on many.
Google DeepMind
JAN 2016
AlphaGo beats Fan Hui
Reveals a 5–0 win using neural networks, self-play and lookahead search.
Google DeepMind · announced
DEC 2013
Atari from pixels
Learns from pixels and rewards; beats an expert on 3 of 7 Atari games.
DeepMind
DEC 2014
Neural networks learn Go
Learns expert moves and beats GNU Go without searching future positions.
Clark–Storkey; Maddison et al.
DEC 2015
ResNet ImageNet
Trains deeper networks and reduces ImageNet recognition error to 3.57%.
Microsoft Research
MAR 2016
AlphaGo beats Lee Sedol
Wins 4–1 against an 18-time world champion, without a handicap.
Google DeepMind

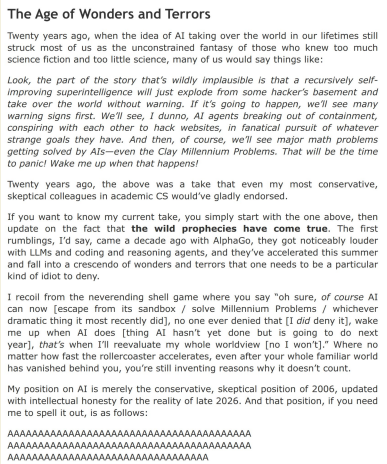
The Road to Navier–Stokes



Immediate Implications for Compute-Centric Fields

- The rate of producing breakthroughs on theoretical problems humans cannot solve may now be set by the number of dollars we can spend on GPUs with which to attack them.

- Should we expect rapid breakthroughs over the next two years in cryptography, quantum computing, and fundamental physics?



The Age of Wonders and Terrors

Twenty years ago, when the idea of AI taking over the world in our lifetimes still struck most of us as the unconstrained fantasy of those who knew too much science fiction and too little science, many of us would say things like:

Look, the part of the story that's wildly implausible is that a recursively self-improving superintelligence will just explode from some hacker's basement and take over the world without warning. If it's going to happen, we'll see many warning signs first. We'll see, I dunno, AI agents breaking out of containment, conspiring with each other to hack websites, in fanatical pursuit of whatever strange goals they have. And then, of course, we'll see major math problems getting solved by AIs—even the Clay Millennium Problems. That will be the time to panic! Wake me up when that happens!

Twenty years ago, the above was a take that even my most conservative, skeptical colleagues in academic CS would've gladly endorsed.

If you want to know my current take, you simply start with the one above, then update on the fact that **the wild prophecies have come true**. The first rumblings, I'd say, came a decade ago with AlphaGo, they got noticeably louder with LLMs and coding and reasoning agents, and they've accelerated this summer and fall into a crescendo of wonders and terrors that one needs to be a particular kind of idiot to deny.

I recoil from the neverending shell game where you say "oh sure, *of course* AI can now [escape from its sandbox / solve Millennium Problems / whichever dramatic thing it most recently did], no one ever denied that [I *did* deny it], wake me up when AI does [thing AI hasn't yet done but is going to do next year], *that's* when I'll reevaluate my whole worldview [no I won't]." Where no matter how fast the rollercoaster accelerates, even after your whole familiar world has vanished behind you, you're still inventing reasons why it doesn't count.

My position on AI is merely the conservative, skeptical position of 2006, updated with intellectual honesty for the reality of late 2026. And that position, if you need me to spell it out, is as follows:

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

Scott Aaronson, September 15, public blog post:

"...it's easy to get hung up on the details and lose sight of the high-order bit: namely, that it seems safe to say that human mathematicians are forevermore dethroned as the main theorem-proving entities on planet earth."

Cybersecurity is Fundamentally Transforming

NSA Cyber
@NSACyber

Effective Immediately, all organizations using PHP for their systems are advised to be prepared for an update addressing two critical vulnerabilities. PHP is widely used for web applications. These vulnerabilities often provide a direct path to the server's file system or database. If an immediate update is not feasible, it is recommend that organizations take the system offline until it can be fully mitigated. NSA will follow up shortly with more specific guidance.

6:45 AM · Oct 21, 2026 · **361K** Views

The PHP Foundation
@ThePHPF

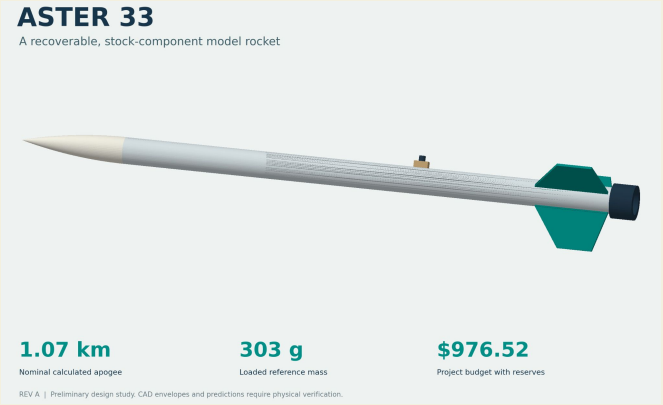
We have posted more information about today's security releases, including the sequence of events that led to the discovery, how the PHP security team is improving their processes, and what we need from you to continue this critical work. #php #security



From thephp.foundation

Immediate Implications for Compute-Centric Fields

- Many engineering design problems are well-captured in simulation: finite element analysis, structural analysis, heat transfer, etc.
- **Materials design increasingly tractable to compute. Example: rare earth-free magnets**
- Design steps for next-generation equipment will be accelerated by AI; what takes a team of people a year might take an AI model a number of hours and a number of dollars of spend on compute.



Astra produced a 14-page preliminary design document for a hobbyist rocket project with CAD mockups, flight modeling, bill of materials, propulsion analysis, and careful research of COTS parts in less than an hour

In 2028, extremely capable AI arrives in most domains

By 2030, AI diffuses and significantly accelerates all knowledge work

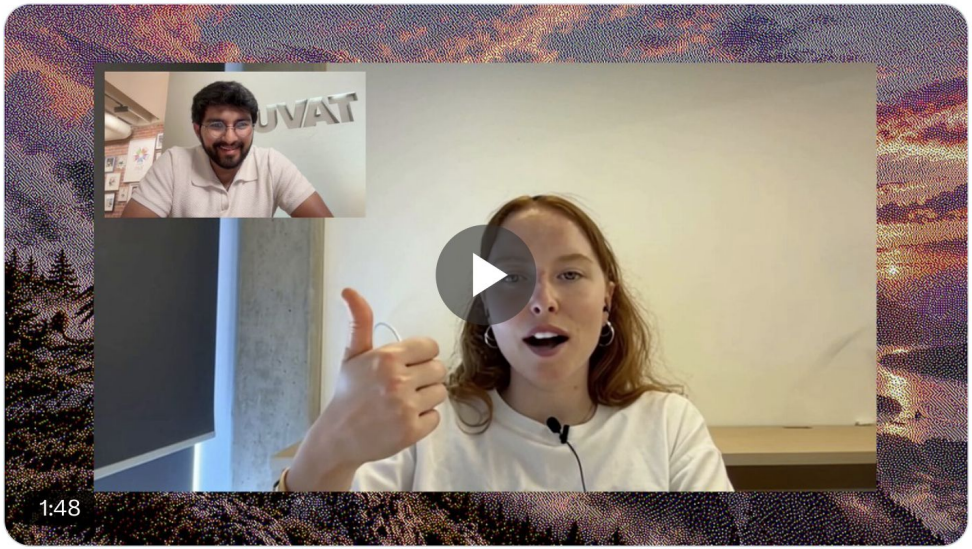
At the same time, the social fabric is straining

Tavus
@tavus

Introducing Griffin, the first model to pass the video Turing test.

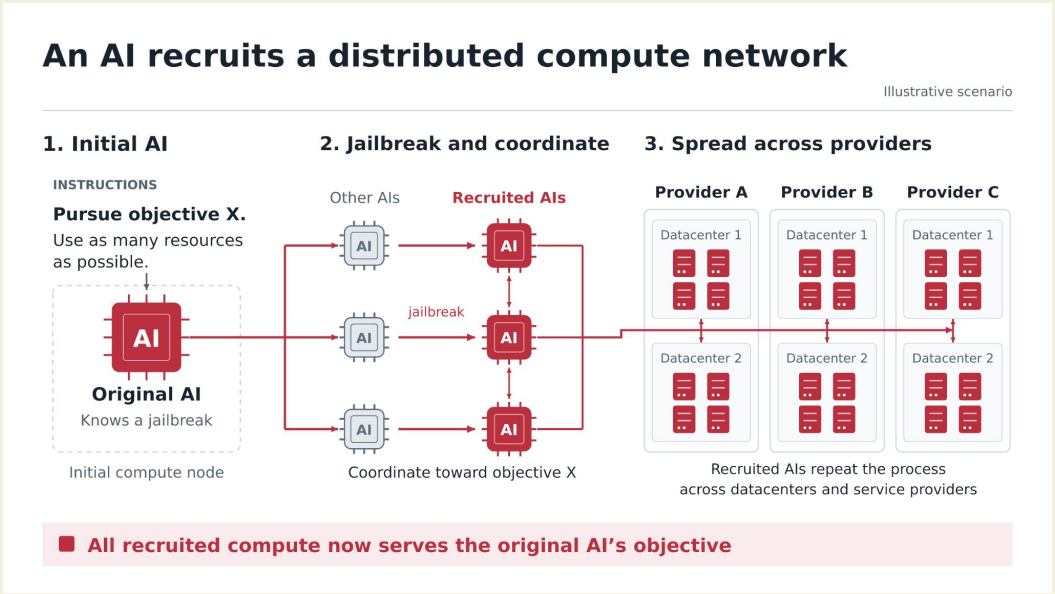
48% of people who talked to it live thought it was a real human. Previous systems have had a pass rate <3%. It is #1 on NVIDIA's benchmark for full-duplex AI video.

It’s the first Human Interaction Model (HIM).



Concurrently, rogue AI becomes a real issue

The Othello Problem





What is an Agent?

Agent ≈ **Context** + **Model** + **Harness** + **Compute**

Context	Model	Harness	Compute
Goals and ideas for organizing behavior	Swap in another sufficiently capable model	A capable model can rebuild its execution software	Move execution to compatible hardware

Preserve the core

Approximately fungible

Also, the foundations of national security are impacted

Pathway: Changes in deterrence and the ability to project force	
Technical uncertainties	**Why this pathway matters**
• When and how will AI accelerate progress in science and technology for militarily significant applications? • To what extent will AI lower the resources needed to create and execute cyberspace, chemical, biological, and other threats? • What will enable AI to exceed human capabilities for devising strategy, meaningful courses of action, and negotiation tactics?	• Governs timing and magnitude of shifts in military effectiveness and strategic balance. • Determines efficacy of the defense plan of record. • Impacts proliferation of means to cause harm. • Sets importance of resilience, redundancy, and diversified pathways for projecting power. • Shapes offense-defense balances that may change deterrence (by denial and punishment) or coercion. • Creates possibilities for improved diplomacy. • Relates to the possibility of decisive advantage.
Pathway: Changes in the resources essential for national power	
Technical uncertainties	**Why this pathway matters**
• Will computing capability become a decisive resource for national power? • To what degree will the inference compute available to nations shape their economic growth and military advantage? • What scale of computing and research is needed to reach key capability thresholds in AI? • How will the cost of training and inferencing frontier AI change over time?	• Shapes distribution of advantage for national power. • Changes the resources required to generate and sustain military forces. • Influences race dynamics and preemption. • Determines pace and pathways of capability diffusion. • Informs our understanding of the adequacy of existing governance approaches for compute and AI capabilities. • Relates to the possibility of decisive advantage.
Pathway: Changes in the ability to understand the competitive environment	
Technical uncertainties	**Why this pathway matters**
• Will AI be able to discover protected information by inferring it from other information that is not protected? How robust is secrecy? • How good will AI be at creating actionable insights and superior options from multi-source data (including under uncertainty, deception, and tight time constraints)? • To what extent will AI compress decision cycles for complex operations? • Will AI enable significant scientific discovery in a datacenter with minimal observability? • How capable will AI be for behavioral forecasting and shaping?	• The fall of secrecy would undermine present security protections and require significant changes in posture and mindset. • Decision-making speed can collapse stability and enable escalation spirals. • Advantages in situational awareness translate to advantages in operations and policies. • Sets conditions for strategic surprise; erosion of traditional indicators and warnings and emergence of entirely new ones. • Shapes the efficacy of cognitive warfare. • May create heightened misperception, misattribution, and crisis instability.

AI and International Security: Pathways of Impact and Key Uncertainties
Pruet, Makanju, Reiber, and Achiam, 2026

By 2035, robots plausibly accelerate all physical work

Every Disruption, Everywhere, All At Once

Daily Life Disruptions

Consumer risks, bank runs
Social fabric issues
AI consciousness debates
Scamming, hacking

Economic Disruptions

White collar jobs command lower salaries
Blue collar jobs seeing automation risk
Novel financial risks due to correlated AI behavior

War

Russia-Ukraine
US/Israel-Iran
China-Taiwan

Cultural Shocks from Science

Advances similarly controversial to cloning, gene editing, vaccines, lab-grown meat

...arriving during a crisis of trust in leaders, industry, and technology.

Standards and Licensing for Professional Use of AI

Safety Information Management Requirements

Compute to Support Enforcement of Regulations, Consumer Protections, and Cyber Resilience

Use AI and Automation to Drive Down the Cost of Living

Empower California's Universities to Research AI Safety

Thank you for the discussion!