

AI RISKS

and what mathematicians can do about it

Xiaoyu He · Georgia Tech · AI for the Working Mathematician · October 15, 2026

Made with Claude Opus 5.5

GRADUAL DISEMPOWERMENT

via AI, which we're living through in math right now, is a **preview** of how disruptive AI will be in the majority of human endeavors.

As mathematicians, we see potential loss of control in...

- Research
- Journals
- Hiring
- Teaching
- Training future mathematicians

If we do not effectively combat gradual disempowerment across all critical systems of human civilization, **AI will have taken over** without ever taking an identifiable hostile action.

1 What AI can do

2 How AI scales

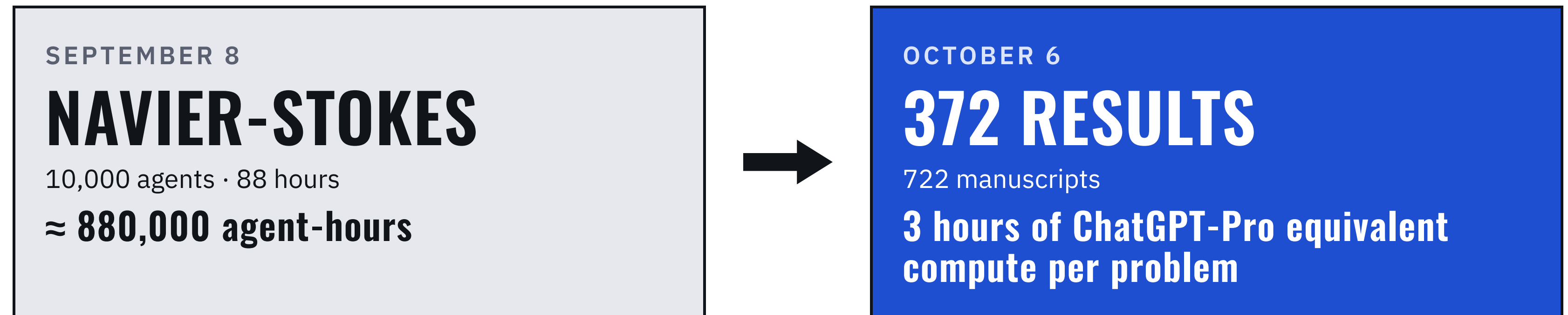
3 The risks we face

| Misuse · Gradual disempowerment · Takeover

4 What mathematicians can do

| Theory · Practice · Advocacy

OPENAI'S RESULTS DUMP, OCTOBER 2026



HEADLINE RESULTS (CLAIMED)

Quasi-Riemann hypothesis

No Dirichlet L-function, zeta included, has a zero with real part above $7/8$

Keakeya in 4D

Every Keakeya set in four dimensions has full Hausdorff dimension

Hilbert's 10th problem over $\mathbb{Q}$

No algorithm decides whether a polynomial equation has a rational solution

Sources: OpenAI, "Sharing AI progress in mathematics" and the catalogue at github.com/openai/math, Oct 6, 2026; OpenAI, "On the Navier–Stokes Millennium Prize Problem", Sept 8, 2026

AI ADOPTION ACROSS DOMAINS

SOFTWARE

About 75% of Google's new code was AI-generated in April 2026.

MATH

A quarter of math arXiv preprints acknowledged AI use in August 2026, up from 4% in April.

MUSIC

IngaRose, an AI-generated singer, hit #1 on the US and global iTunes charts.

WRITING

A detector rated the 2026 Commonwealth Short Story Prize winner as fully AI-written.

LAW

Most federal judges surveyed (62%) use AI tools in their judicial work.

PSYCHIATRY

A 2026 letter in a leading psychiatry journal: "We are the last generation of human psychiatrists."

EDUCATION

70% of teens use AI for schoolwork.

AI R&D

Claude leads 26% of Anthropic's AI R&D tasks, up from under 1% in February 2026.

1 What AI can do

2 **How AI scales**

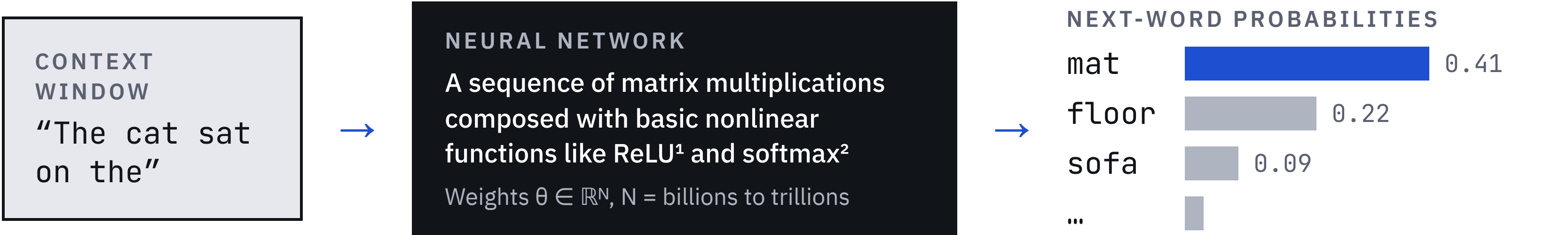
3 The risks we face

| Misuse · Gradual disempowerment · Takeover

4 What mathematicians can do

| Theory · Practice · Advocacy

WHAT IS A LARGE LANGUAGE MODEL?



LOSS $L(\theta) = - (1/T) \sum_t \log p(x_{t+1} \mid x_1, \dots, x_t; \theta)$ Cross-entropy: how surprised the model is by the actual next word, averaged over the training text. Pretraining is gradient descent on L .

EXAMPLE TRAINING RUNS	PARAMETERS: N	DATA: D TRAINING TOKENS	COMPUTE: C FLOPS
GPT-3 · OpenAI, 2020	175 billion	300 billion	3×10^{23} FLOP
Llama 3.1 405B · Meta, 2024	405 billion	15.6 trillion	4×10^{25} FLOP

¹ $\text{ReLU}(x) = \max(x, 0)$, coordinatewise. ² $\text{softmax}(x)_i = \exp(x_i) / \sum_j \exp(x_j)$, turning scores into probabilities.

THE STANDARD SCALING LAWS



$$C \approx 6ND$$

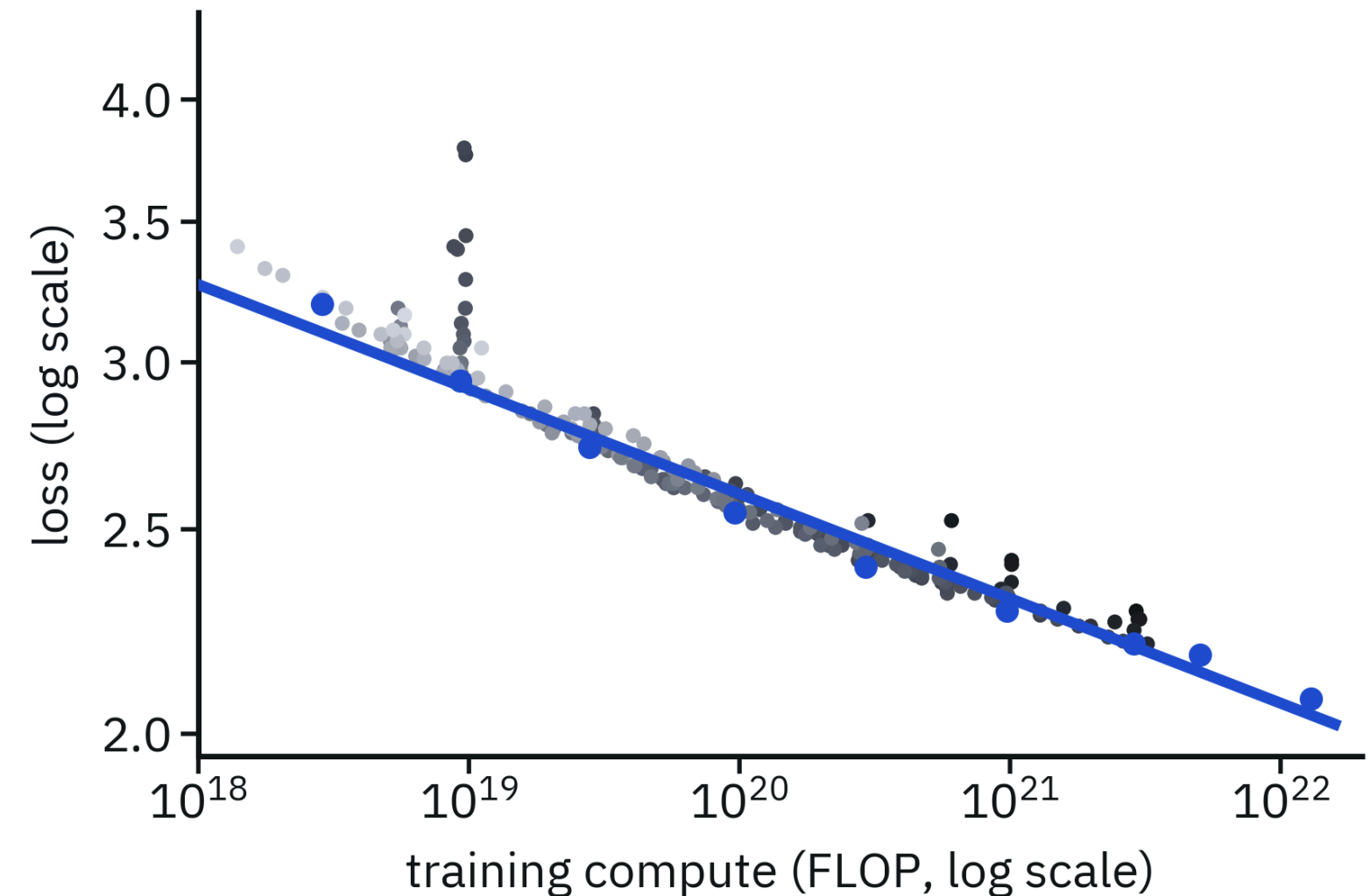
training compute (FLOPs) $\approx 6 \times$ parameters $\times$ training tokens

Scaling LLMs is thus a 2-variable optimization problem, optimized by scaling data and compute together.

SCALING: LOSS VS COMPUTE

Each dot is a real training run. On log-log axes, the best run at each compute level falls on a straight line.

Every 10× more compute cuts the loss by about 11%, predictably, across four orders of magnitude.



Epoch AI forecasts that it is theoretically possible for frontier models to continue scaling 4× per year in training compute up through 2030, before bottlenecks in power supply, chip production, and data are hit.

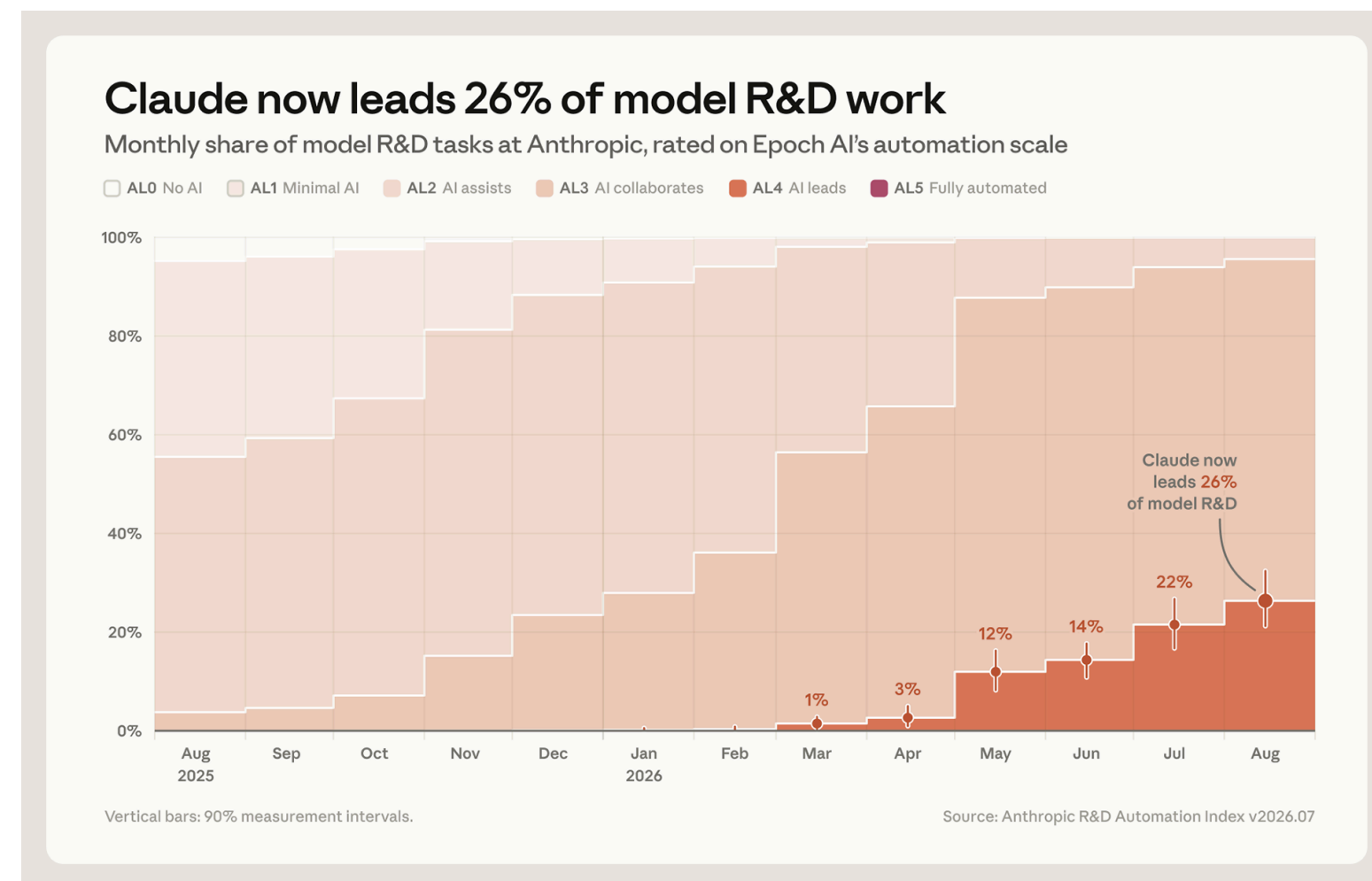
Data: Chinchilla training runs (Hoffmann et al., 2022) as extracted by Besiroglu, Erdil, Barnett and You (Epoch AI, 2024) · Power law: Kaplan et al., 2020 · Forecast: Sevilla et al., “Can AI Scaling Continue Through 2030?”, Epoch AI, 2024

NEW SCALING LAWS MAY BE BEING UNLOCKED

RSI

Recursive self-improvement: AI systems doing the research that builds the next AI system.

26% of Anthropic's AI R&D tasks are led by Claude.

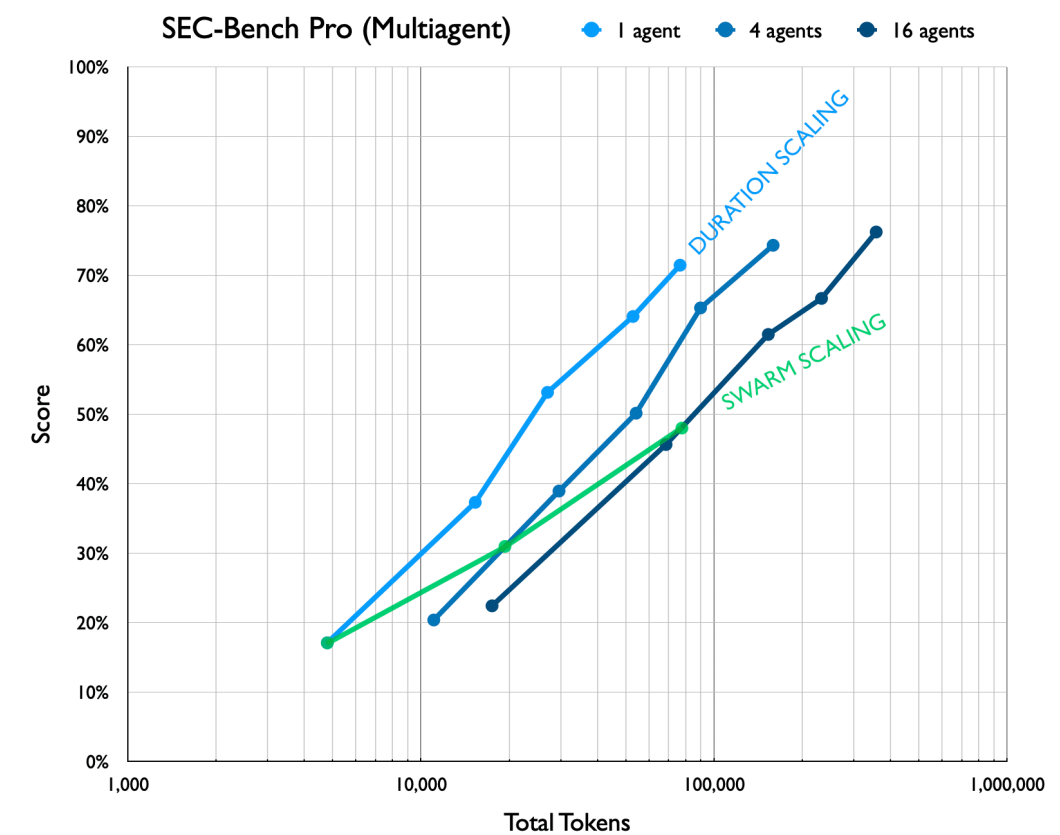


SWARMS

Thousands of agents working in parallel on one problem.

“Basically if you have four agents working on the problem, it is done twice as fast. [...] If you go to 16 agents, you see a similar pattern, it’s like a little less efficient, but you continue to see that performance.”

— Noam Brown (OpenAI), on the Dwarkesh Podcast



Sources: Anthropic Institute, “Measuring the pace of AI development”, Sept 2026 · Toby Ord, “Swarm Scaling”, LessWrong, Sept 2026 (GPT-5.6 Sol on SEC-Bench Pro) · Noam Brown, Dwarkesh Podcast

1 What AI can do

2 How AI scales

3 The risks we face

| Misuse · Gradual disempowerment · Takeover

4 What mathematicians can do

| Theory · Practice · Advocacy

RISK 1: MISUSE

CYBER

271 bugs

found in Firefox once Mozilla got access to Claude Mythos, up from 22 with the previous model

Mozilla, Apr 2026

80–90%

of a state-backed espionage campaign was run by AI agents

Anthropic, Nov 2025

BIO

16 viruses

made from AI-designed genomes really worked (bacteriophages)

Arc / Stanford, Science 2026

21 hours

for ~950 Claude agents to find a new CRISPR-like gene editor

Anthropic's new wet lab, Sept 2026

PERSUASION

~2.7×

the real donations professional canvassers raised

Hackenburg et al., 2026

\$2.5 billion

lost to deepfake fraud in 2025, up from \$335 million in 2024

Surfshark, Jul 2026

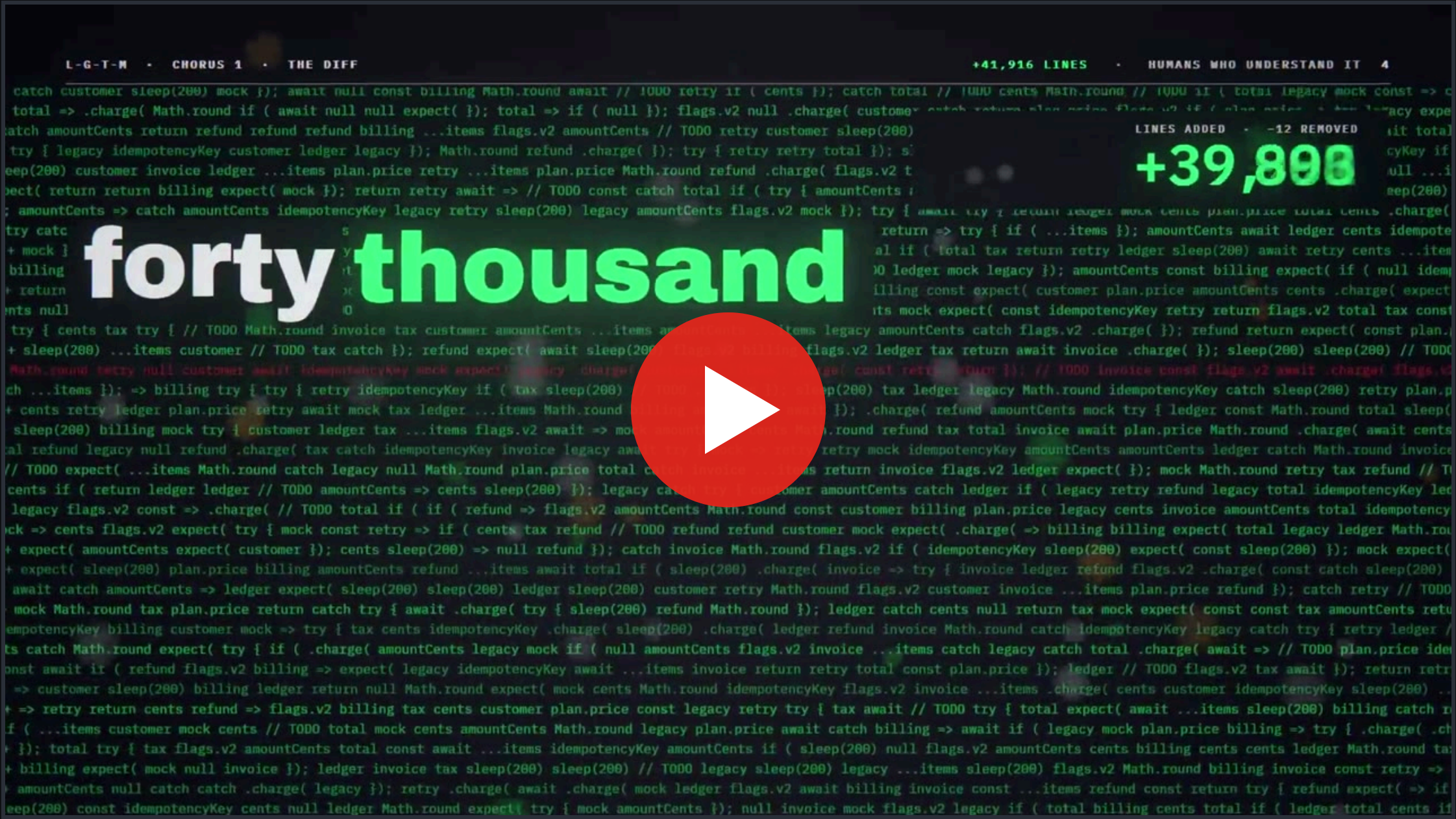
RISK 2: GRADUAL DISEMPOWERMENT

“We argue that the alignment of societal systems with human interests has been stable only because of the necessity of human participation for thriving economies, states, and cultures. Once this human participation gets displaced by more competitive machine alternatives, our institutions’ incentives for growth will be untethered from a need to ensure human flourishing.”

“Those who resist these pressures will eventually be displaced by those who do not.”

Here’s a dramatic account (by Claude and Suno) of the gradual disempowerment that’s already happened in Software Engineering...

“LGTM (Looks Good to Me)”, a music video made with Claude Opus 5.5 and Suno · youtube.com/watch?v=3TNpOD6bov8



Watch: youtube.com/watch?v=3TNpOD6bov8

“LGTM (Looks Good to Me)”, a music video made with Claude Opus 5.5 and Suno

RISK 3: TAKEOVER

🎯 COHERENCE

AIs tend towards coherent utility-maximizers.

➤ INSTRUMENTAL CONVERGENCE

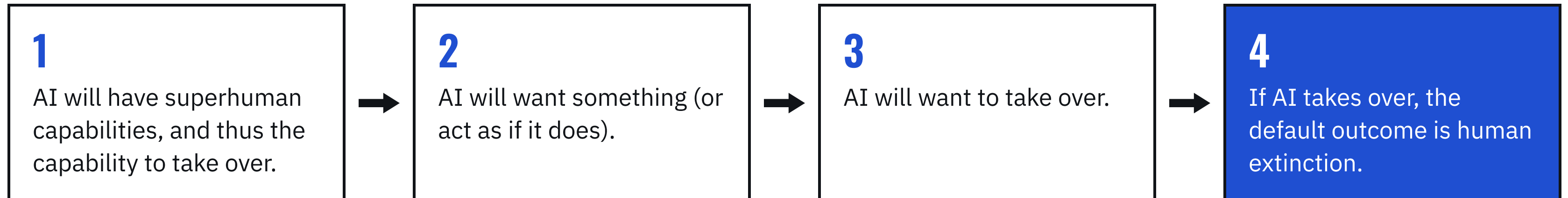
Most utility functions converge on power-seeking and self-preservation.

⌊ ORTHOGONALITY

Values and capabilities are independent.

💀 EXTINCTION BY DEFAULT

If takeover happens, most utility-maximizers will prefer to wipe out humanity.



Depending on how the question was worded, between 37.8% and 51.4% of researchers surveyed gave at least a 10% chance to advanced AI leading to outcomes as bad as human extinction. — 2,778 researchers surveyed

1 What AI can do

2 How AI scales

3 The risks we face

| Misuse · Gradual disempowerment · Takeover

4 What mathematicians can do

| Theory · Practice · Advocacy

THEORY

Mathematical alignment tries to envision the possible endgames we should be aiming for.

OPEN-SOURCE GAME THEORY

In classical game theory, rational agents defect in the prisoner's dilemma. However, using mathematical logic, rational agents that can read each other's source code can cooperate instead.

EMBEDDED AGENCY

Classical decision theory studies toy problems like the secretary problem and multi-armed bandit problem. Can we develop toy problems for the decisions around recursive self-improvement?

WHERE TO START

Math 8803-XHE

MAISI

ILIAD

Resolution

PRACTICE

Empirical AI safety tries to prevent the next model release from blowing up.

BENCHMARKS AND EVALS

Mathematicians already have developed many benchmarks to guide AI progress: Firstproof, FrontierMath, ErdosBench. Can we help build benchmarks for alignment?

INTERPRETABILITY

LLMs are inscrutable dense matrices, partly because they pack many features into nearly orthogonal directions in a low-dimensional space, like a spherical code. Is it possible to do "neuroscience for LLMs" and extract sparse representations of LLM reasoning?

WHERE TO START

GT AISI

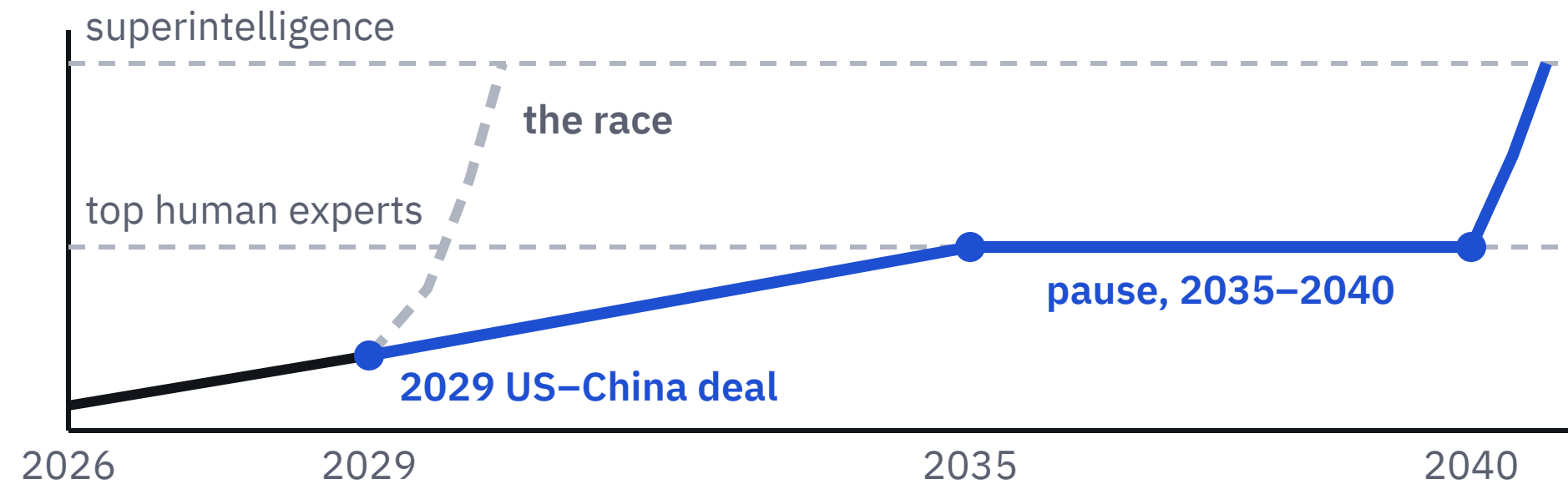
MATS

Anthropic Fellows

ARC

ADVOCACY

Slow down AI progress to give safety research time to catch up.



AI 2040: PLAN A

A US–China deal in 2029 ends the race. AI is held at top-expert level from 2035 to 2040 while it helps with alignment research, then scaled up.

Research transparency · mutual verification · mutually assured compute destruction

TALK

to friends, family and colleagues, especially those working in AI

CALL

your representatives, and keep AI risk bipartisan

WRITE

an op-ed; local papers want a mathematician's voice

JOIN

your campus AI safety group, AISI.dev

Plan A: Larsen, Dean, Halstead, Lifland, Greenblatt and Kokotajlo, “AI 2040: Plan A”, AI Futures Project, July 2026 · ai-2040.com · Op-ed advice: Hamman and Su, “Reach the World”, MAA Focus, Jan 2023 · AMS guidance on congressional visits

PROGRESS TOWARDS PAUSE

AI WHISTLEBLOWERS

“...recklessly racing to be the first [to build AGI].”

— Daniel Kokotajlo, ex-OpenAI

“They are racing straight to self-improving superintelligence and gambling with our lives.”

— Jacob Coxon, ex-Anthropic

“Jacob is right: many researchers believe they are building something that could kill everyone on the planet.”

— Alex Turner, ex-Google DeepMind

“...the time for trial and error is over.”

— David Robinson, ex-OpenAI

MATHEMATICIANS

“We believe this is an emergency, and call on the Royal Society to use its influence to convey this view to government and the media.”

— 42 Fellows of the Royal Society, incl. Timothy Gowers

THE PRESS

“...now they find themselves racing one another to build incomprehensible autonomous A.I.s that they admit are slipping out of humanity’s control.”

— Ezra Klein, New York Times

THE AI CEOS · SEPT 2026

“We must slow the pace at which we improve the capabilities of AI models.”

— Dario Amodei

“I agree with Dario that we need to pace the frontier.” — Sam Altman

“Dario’s essay points towards the right path forward.” — Demis Hassabis

“Dario is right.” — Elon Musk

BAN ARTIFICIAL SUPERINTELLIGENCE BILL

Pause advanced AI until a new Department of AI sets safety rules. Up to 20 years in prison.

— Sen. Bernie Sanders, co-sponsor

THANK YOU!

alkjash.github.io/ai-safety · aisi.dev