MKT927: Intro to Quantitative Marketing

Lecture 11: AI, the big picture

Which AI tools do you currently use and for what?

In what year will an AI be fully able to do your research?

Outline

AGI - Situational Awareness

The Macroeconomics View

AI Narrow View - Prediction Policy Problems

AI - Concepts

AGI - Artificial General Intelligence - AI as smart as humans.

ASI - Artificial Super Intelligence - AI substantially smarter than the smartest human.

FLOPs - Floating Point Operations - Used to measure compute capacity.

Training vs Inference (Training model vs having the model produce tokens)

Agent - AI model that can interact with the (digital) world without human intervention.

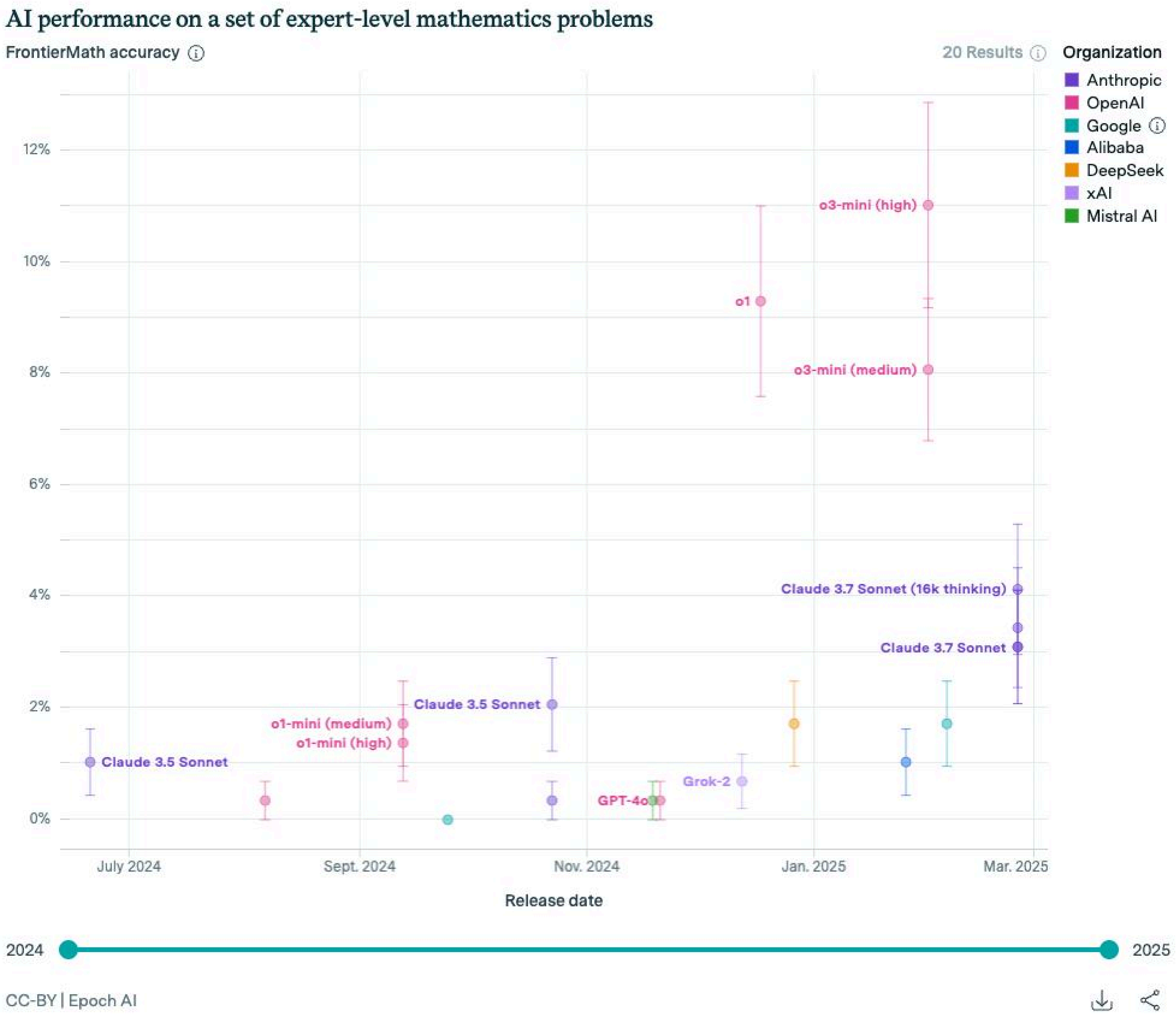
Evaluations / Evals - Benchmarks used to judge how good AI models are at various tasks.

RLHF - Reinforcement learning with human feedback.

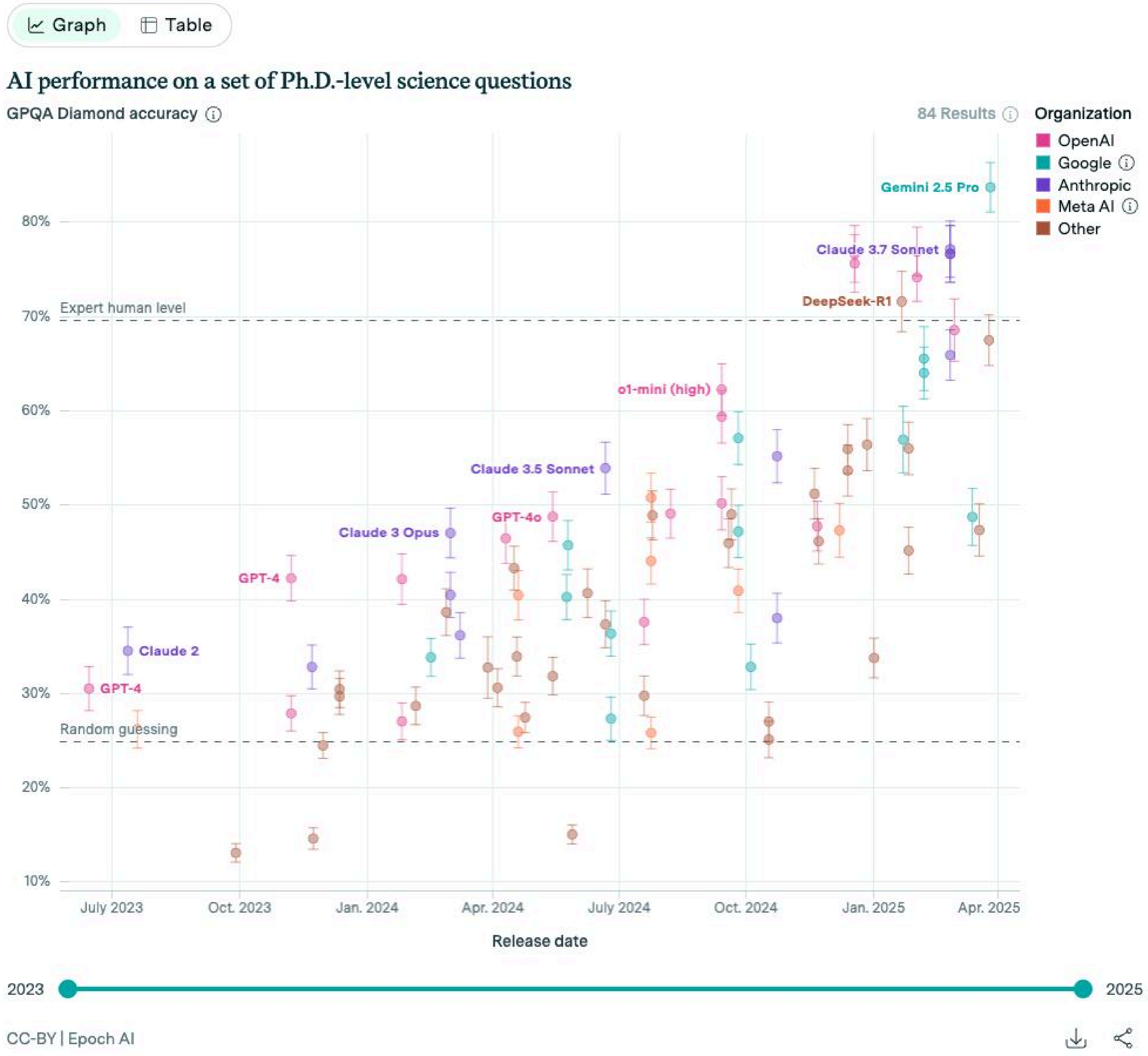
CoT - Chain of thought. Often combined with reinforcement learning.

Reasoning → Using tokens to 'think' through a problem. Although, the tokens may not be the reason for the final answer (see work by Anthropic).

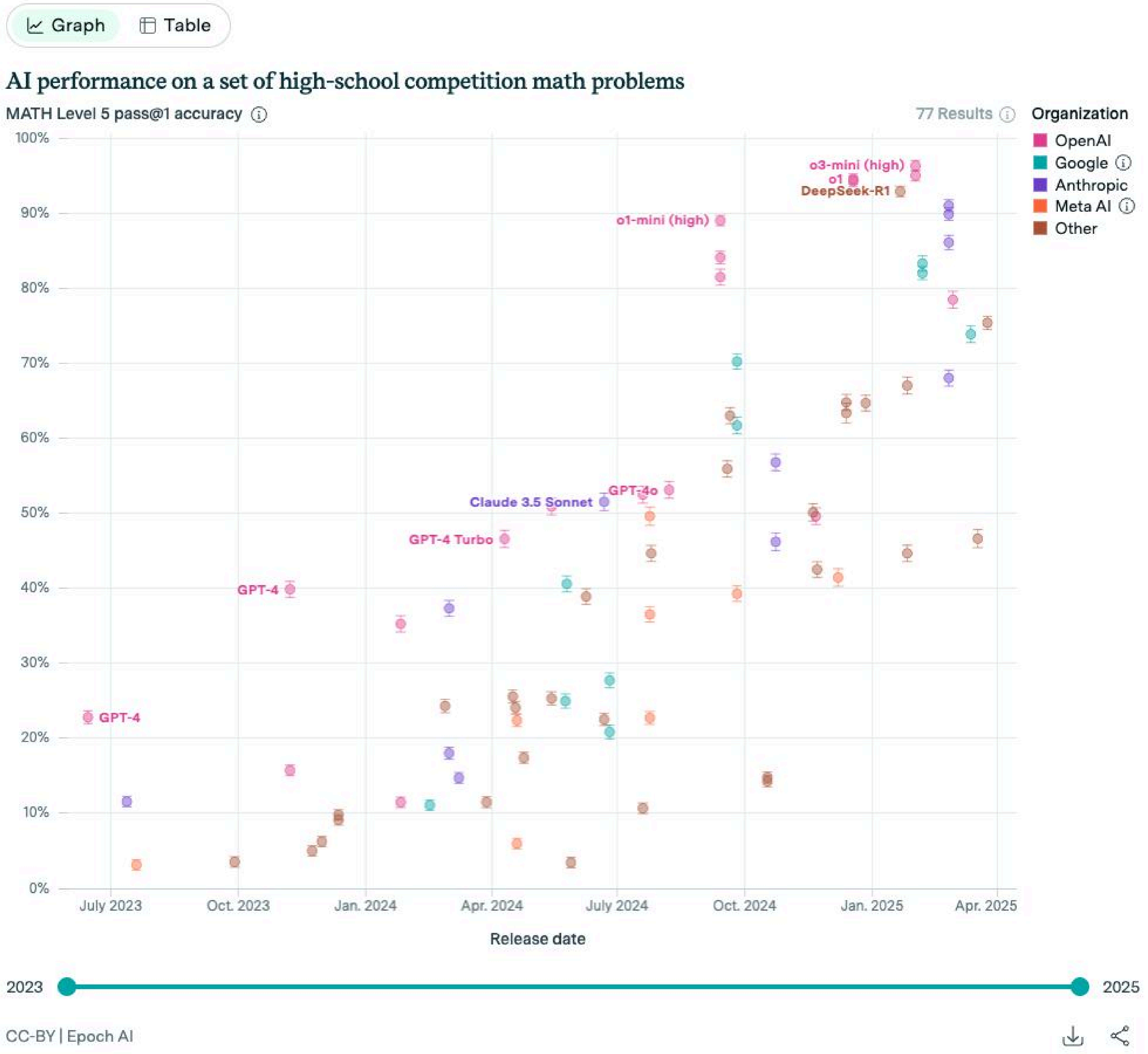
Current AI capabilities



Current AI capabilities



Current AI capabilities



Current AI capabilities

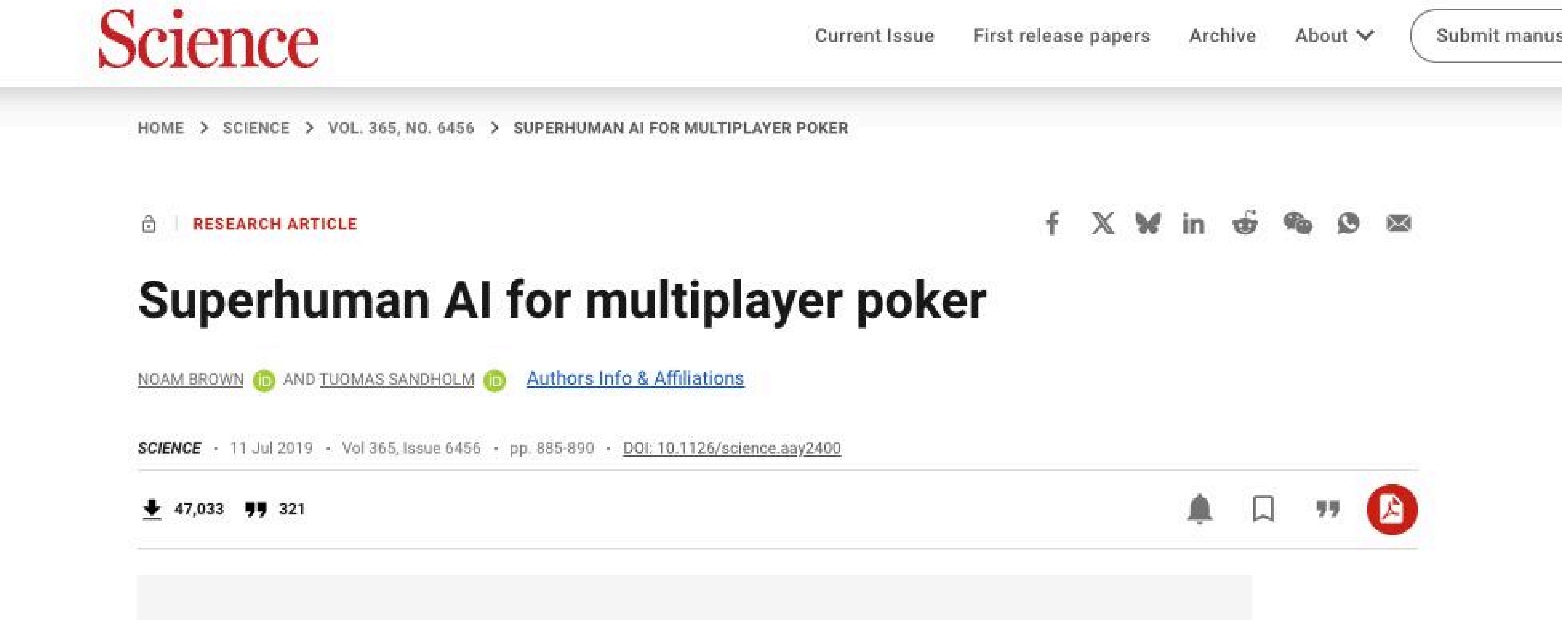
Discovery of protein structure and new molecules.

Navigation of websites - very rapid improvement.

Self-driving cars.

Robotics - still in progress.

Dominance in games such as Chess, Go, and Poker.



Almost everyone underestimated the rate of AI progress

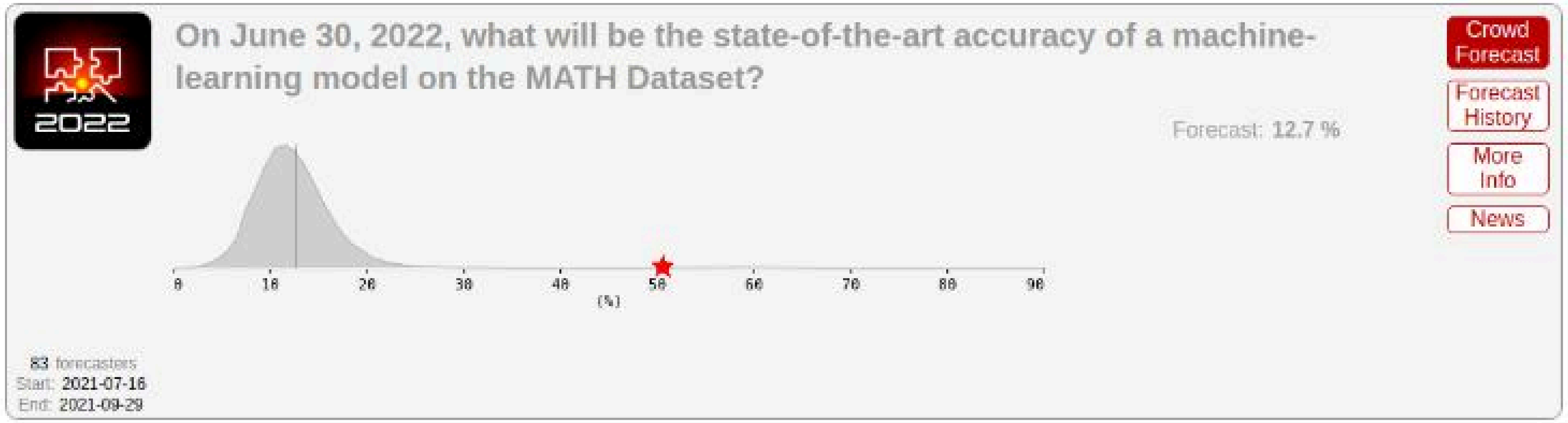


Figure 8: Gray: Professional forecasts, made in August 2021, for June 2022 performance on the MATH benchmark (difficult mathematics problems from high-school math competitions). Red star: actual state-of-the-art performance by June 2022, far exceeding even the upper range forecasters gave. The median ML researcher was even more pessimistic.

Scaling Laws

[PDF] Scaling laws for neural language models [PDF] arxiv.org

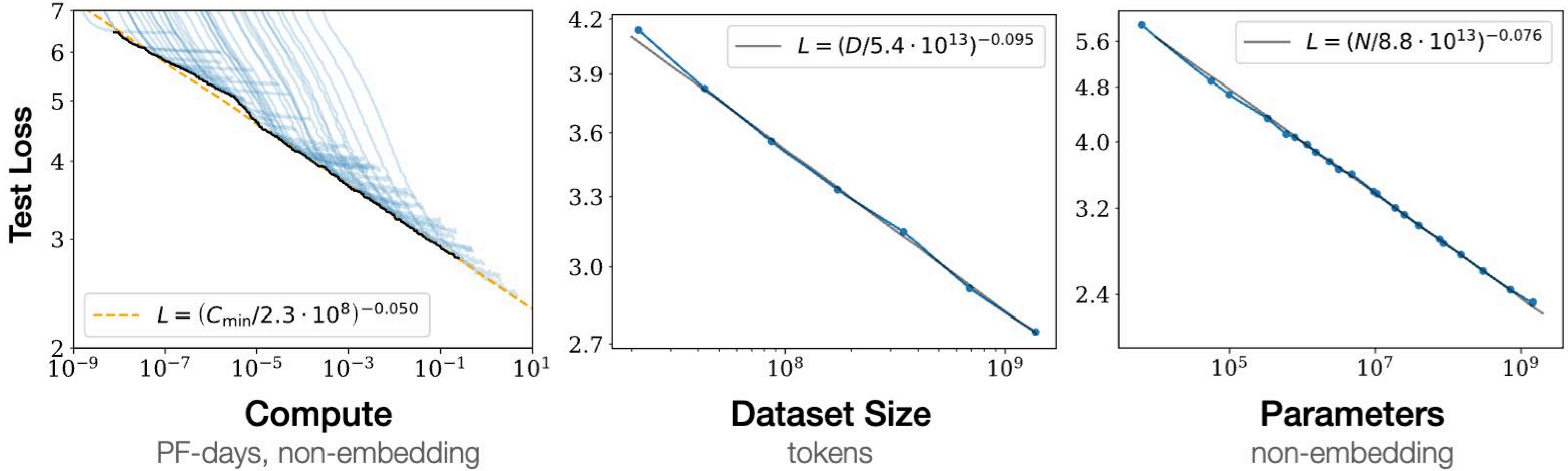
J Kaplan, S McCandlish, T Henighan, TB Brown, B Chess, R Child, S Gray, A Radford, J Wu…

arXiv preprint arXiv:2001.08361, 2020 · arxiv.org

Abstract

We study empirical scaling laws for language model performance on the cross-entropy loss. The loss scales as a power-law with model size, dataset size, and the amount of compute used for training, with some trends spanning more than seven orders of magnitude. Other architectural details such as network width or depth have minimal effects within a wide range. Simple equations govern the dependence of overfitting on model/dataset size and the dependence of training speed on model size. These

SHOW MORE



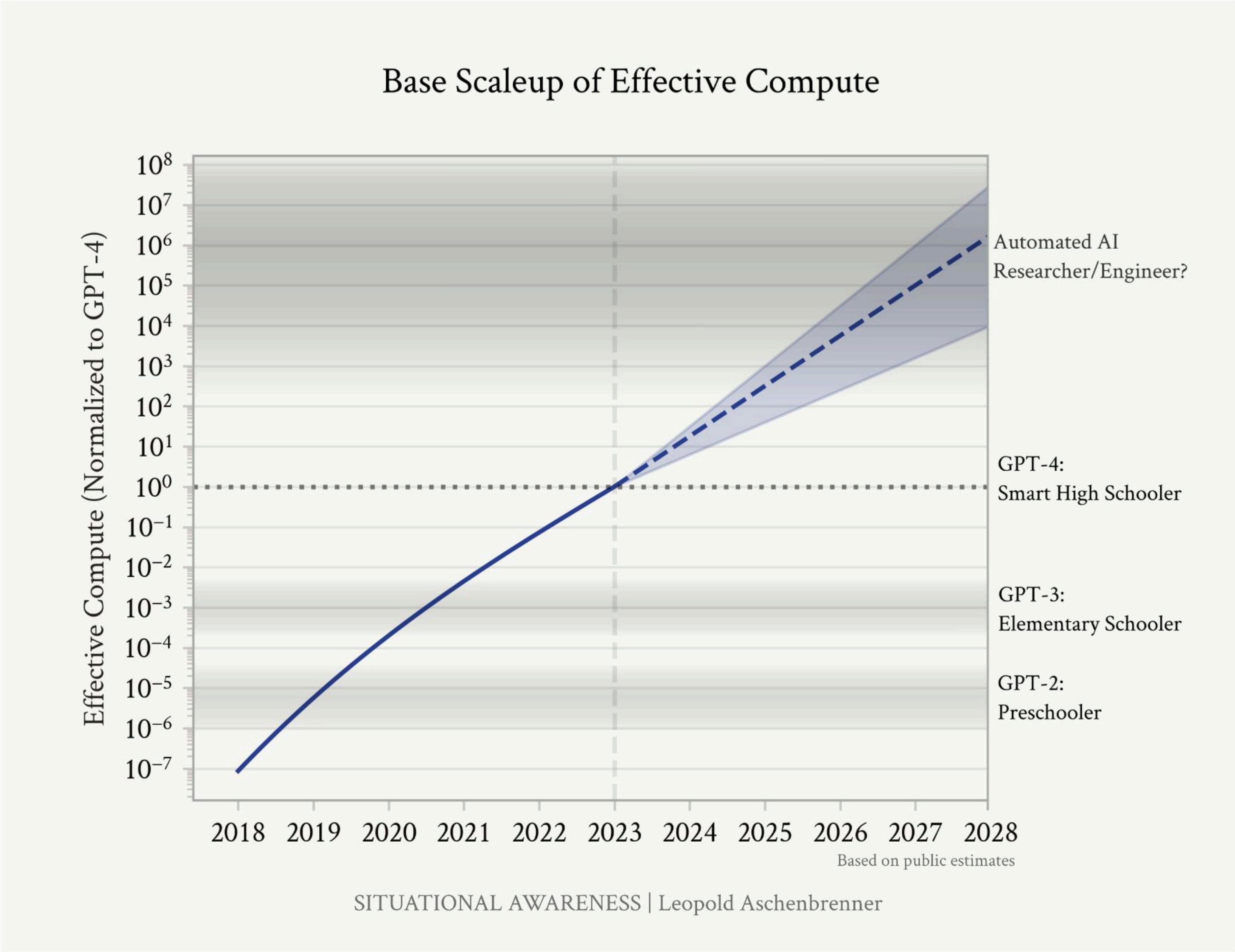
Situational Awareness

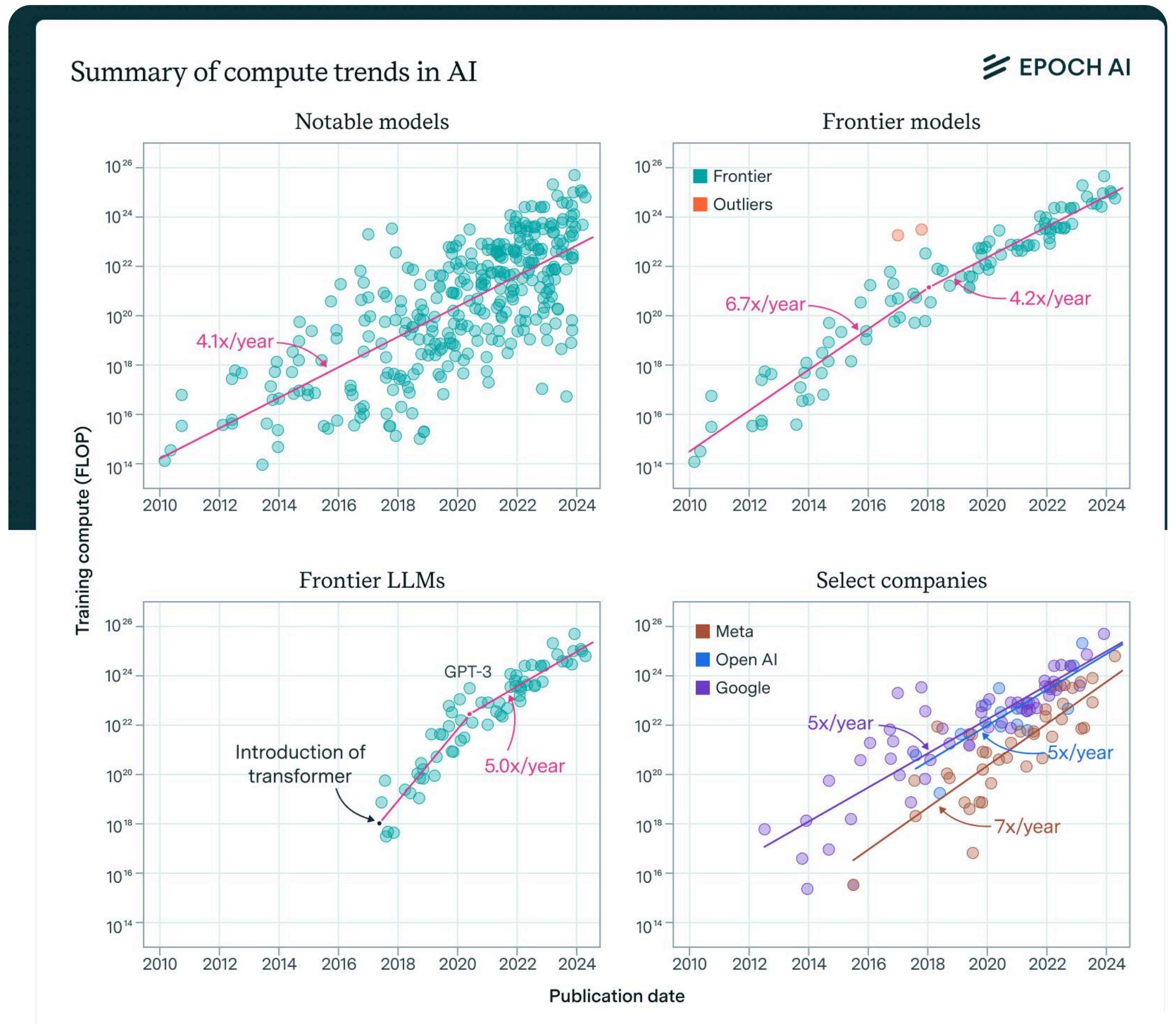
How do we extrapolate from here? Things will keep on improving.

Use effective compute as a benchmark for potential AI capabilities and count on scaling laws continuing.

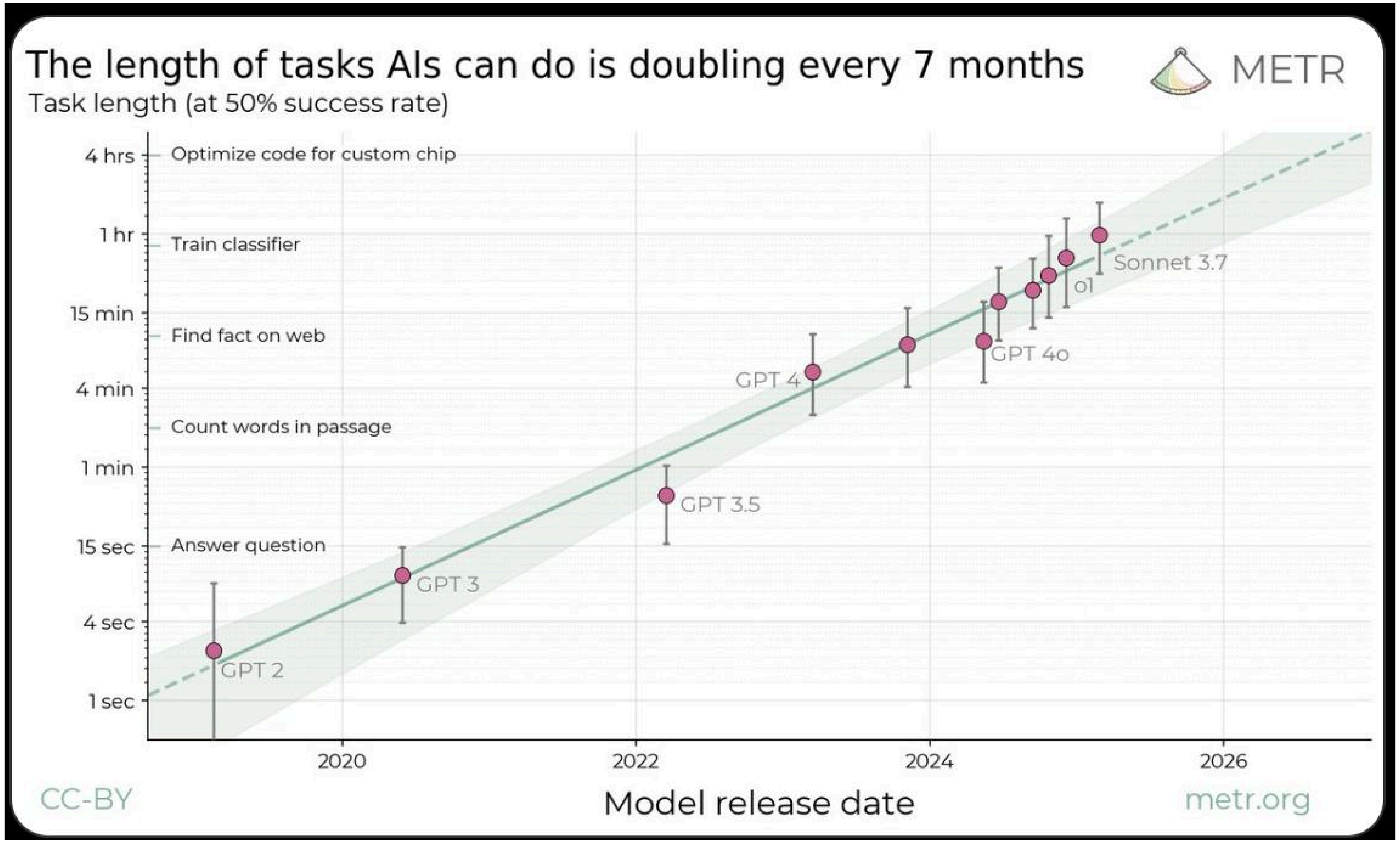
Scaling progress comes from:

- Compute
- Algorithmic efficiency
- “Unhobbling Gains”
- Data





Scaling the duration of work



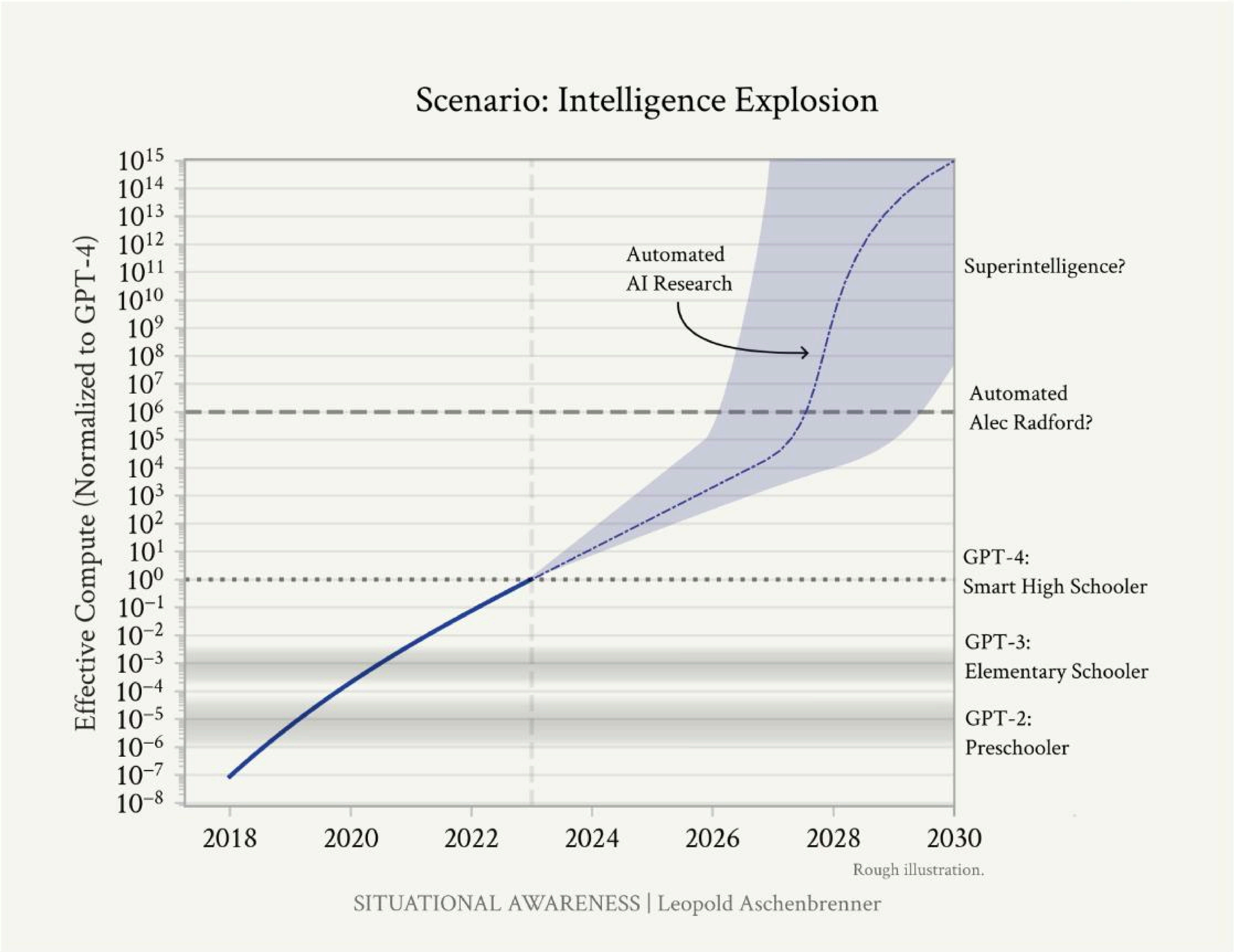
AGI to ASI

Why would intelligence stop at a human level? It hasn't in any specific domain.

Once you have one AGI, would won't you have many AGIs?

Especially important, using AGI to automate AI research in order to produce ASI.

Plausible timelines using naive extrapolation for this to happen point to 2027/2028.

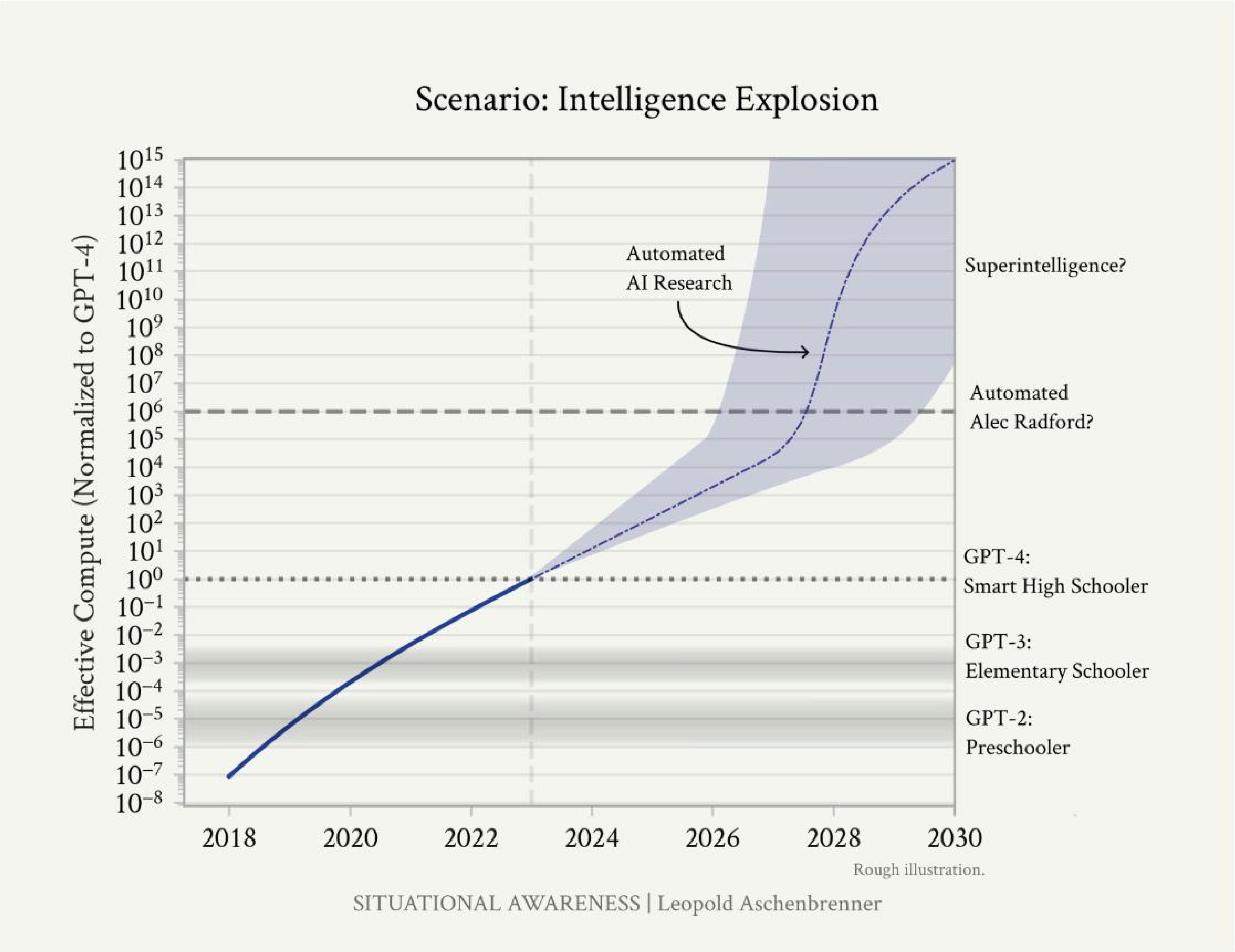


Using AI to automate AI research

- Plausible use case for LLMs. Why? They don't need to interact with the real world.

- We can run millions of copies thinking at speeds much faster than human researchers.

- "100 million automated Alec Radfords"



Implications of ASI

- Science fiction is the best guide.
- Key point: they are qualitatively different from humans.
  - They will do things that make little sense to us but that are correct. Example from AlphaGo, but in every dimension.
- “Obvious” implications:
  - Robotics becomes very useful.
  - Drastic increase in energy availability but also consumption.
  - Acceleration of scientific research.
  - Military advantages.
- ASI but with a will of its own:
  - Can overthrow governments, conquer people, etc…

Second order implications

- If this technology is so transformative / valuable, society will invest in it.
- The scale of the investment will be unprecedented. Leopold’s prediction ($1T in 2027).
- Huge energy requirements: clusters with power requirements of medium sized US states.
- Leopold assumes global economy can provide this level of production based on extrapolations.
- Role for government: investment, regulation, competition between US and China.
- Prediction: “War” footing in the AI race.

Alignment / Safety

- If you believe at least 50% of the above is true, you should be thinking about alignment.
- Alignment: Keeping humans in control of the AIs, even when the AIs get very smart. Note, this is an unsolved problem even with current LLMs.
- Alignment: Preventing AIs from doing something very bad as a side effect of what they were told to do (paperclip maximizers).
- Much bigger version of existing research streams in social science about algorithmic bias, etc…

AI 2027

Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, Romeo Dean

We predict that the impact of superhuman AI over the next decade will be enormous, exceeding that of the Industrial Revolution.

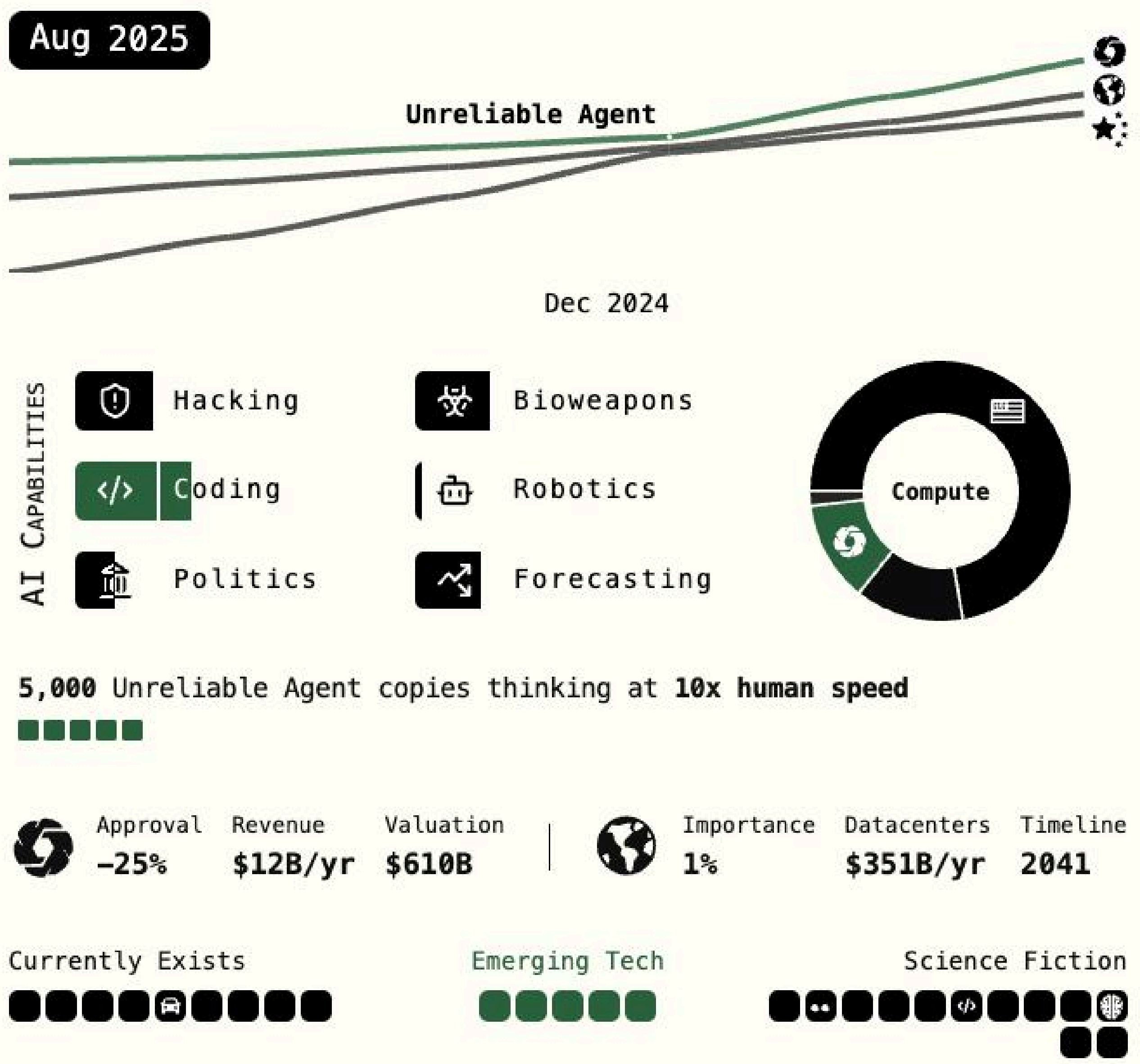
We wrote a scenario that represents our best guess about what that might look like.[1] It's informed by trend extrapolations, wargames, expert feedback, experience at OpenAI, and previous forecasting successes.[2]

What is this? How did we write it? Why is it valuable? Who are we?

Published April 3rd 2025 | PDF | Listen

Mid 2025: Stumbling Agents

The world sees its first glimpse of AI agents.



AI 2027 - The viewpoint

- Recent work in the same theme as Situational Awareness.
- Predictions:
  - Mid 2025: Useful agents that are like personal assistants.
  - Model trained at 10^28 FLOP (3 OOMs more than GPT-4).
  - 2026: Coding automation for large parts of coding. Can already see how this could happen.
  - Late 2026: AI starts “taking” jobs
  - Late 2026: The stock market has gone up 30%.
  - January 2027: Model that if escaped could survive and replicate autonomously.
  - Mid 2027: Self-improving AI + cheap remote worker.

AI Regulation

- Self-regulation: Anthropic's Responsible scaling policy.
- SB1047
  - Controversial bill vetoed by Gavin Newsom.
  - Coverage threshold: 10^26 FLOP or $10M fine tuning.
  - Requirements:
    - Submit for certification.
    - Mitigations for critical harms (bioweapons, cybersecurity, autonomous crimes).
    - Have a kill switch.

In our updated policy, we have refined our methodology for assessing specific capabilities (and their associated risks) and implementing proportional safety and security measures. Our updated framework has two key components:

- **Capability Thresholds:** Specific AI abilities that, if reached, would require stronger safeguards than our current baseline.
- **Required Safeguards:** The specific ASL Standards needed to mitigate risks once a Capability Threshold has been reached.

At present, all of our models operate under ASL-2 Standards, which reflect current industry best practices. Our updated policy defines two key Capability Thresholds that would require upgraded safeguards:

- **Autonomous AI Research and Development:** If a model can independently conduct complex AI research tasks typically requiring human expertise—potentially significantly accelerating AI development in an unpredictable way—we require elevated security standards (potentially ASL-4 or higher standards) and additional safety assurances to avoid a situation where development outpaces our ability to address emerging risks.
- **Chemical, Biological, Radiological, and Nuclear (CBRN) weapons:** If a model can meaningfully assist someone with a basic technical background in creating or deploying CBRN weapons, we require enhanced security and deployment safeguards (ASL-3 standards).

Outline

AGI - Situational Awareness

The Macroeconomics View

AI Narrow View - Prediction Policy Problems

Big picture - Technology drives GDP growth of advanced economies.

- Standard macroeconomic models do not model technology and take it as a residual. The Solow model.
- Key models of technological growth in macroeconomics:
  - Romer: Ideas produced by people result in increases in the productivity of the economy.
  - Weitzman: New ideas come from combinations of old ideas. “Combinatorial” problem.
  - Kremer: The O-ring model, where the worst component of a production process plays a disproportionate role.

General Purpose Technologies

- Disproportionate influence of some technologies on economic progress:
  - Steam Engine
  - Electricity
  - Semi-conductors
- The diffusion of these technologies throughout the economy took a long time.
- In fact, some parts of the world still do not have them.

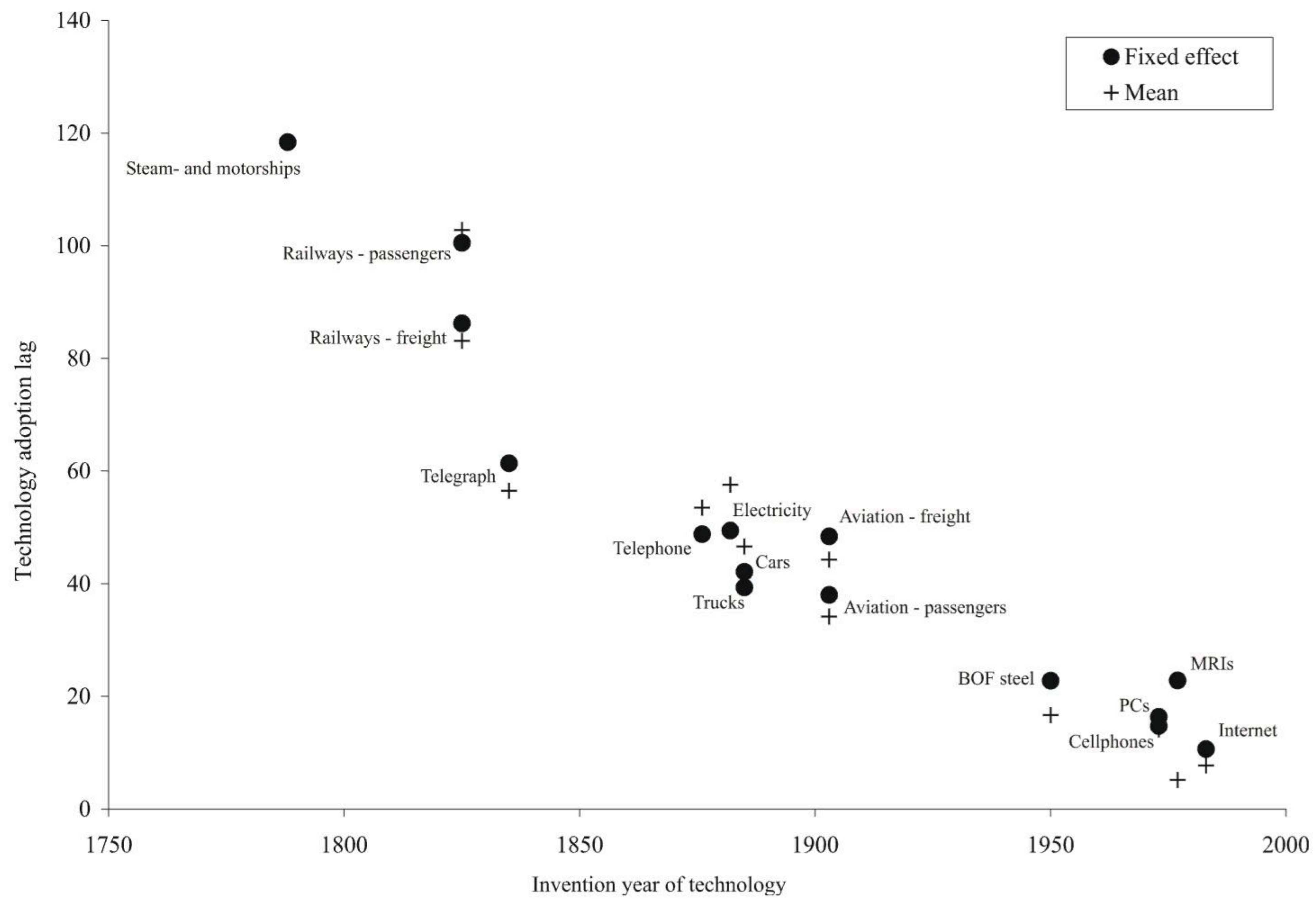
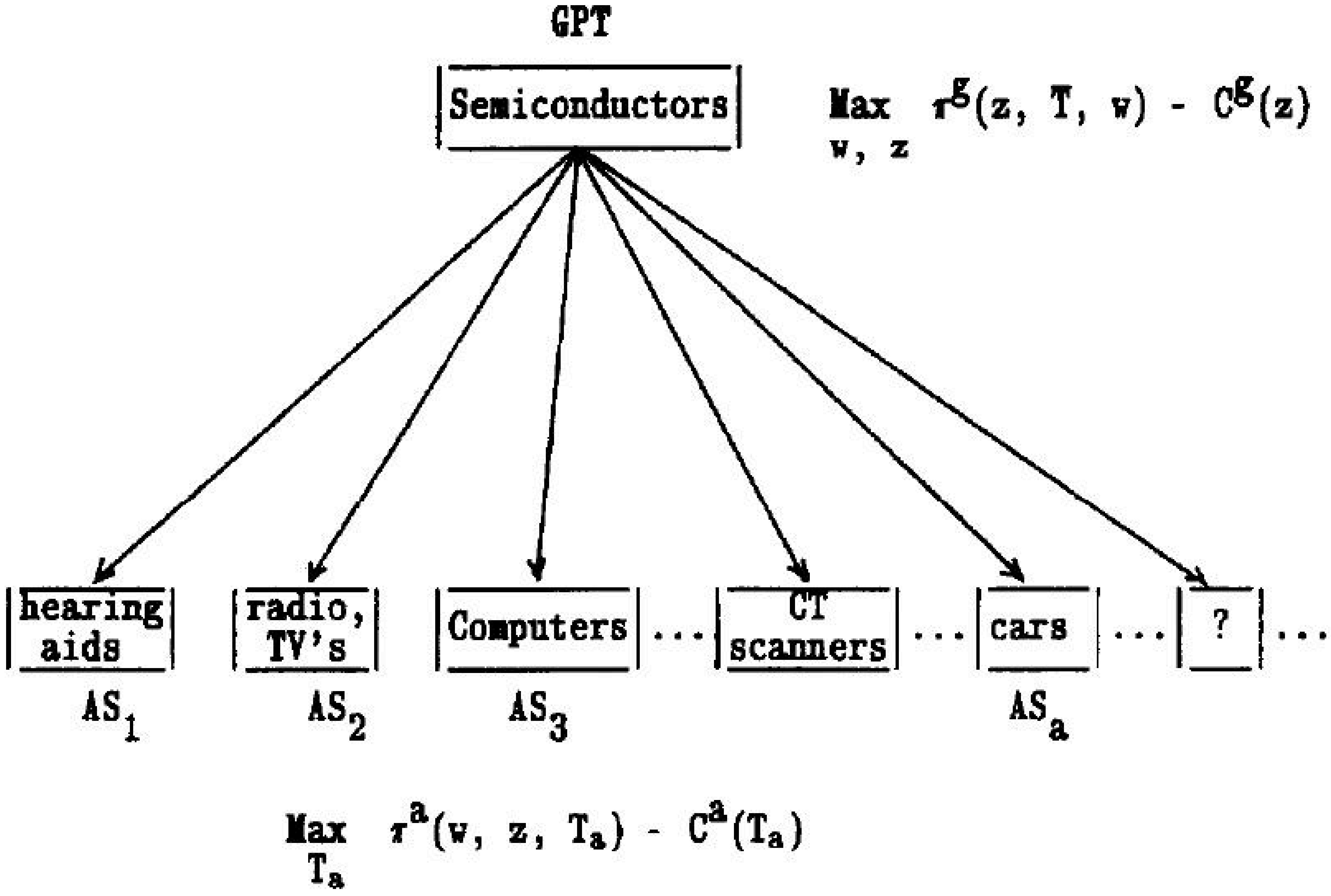


Figure 2: Technology adoption lags decrease for later inventions

What are General Purpose Technologies (GPTs)?

- GPTs are characterized:
  - Pervasiveness: Usable in a lot of sectors.
  - Potential for improvement: They get better over time.
  - Innovational complementarities (IC): Need R&D to apply GPT to specific applications.
- GPTs are enabling technologies - often not useful as an end product but require innovation in applications as well.
  - Electric motors → more efficient factory design
  - Semiconductors → innovative applications in multiple industries (hearing aids, etc…)

Model


GPT
Semiconductors
Max $\pi^g(z, T, w) - C^g(z)$ w, z
hearing aids
radio, TV's
Computers
CT scanners
cars
?
...
AS_1
AS_2
AS_3
AS_a
Max $\pi^a(w, z, T_a) - C^a(T_a)$ T_a

Vertical Implications

GPT – Application Sector (AS) Relationship: Vertical linkage; GPT as an input to AS innovation.

Vertical Externality Explained: Innovation in the GPT benefits AS innovation (IC). But the GPT innovator may not fully capture the returns generated in application sectors.

"Too Little Innovation" in GPT: Monopoly pricing by GPT firms underprovides quality (z) because they don't internalize the full social benefit (including AS surplus).

Dual Appropriability Problem: AS innovation also benefits GPT demand, but AS firms may not fully internalize this feedback loop.

Horizontal Implications:

Horizontal Linkages: Among Application Sectors: Multiple ASs utilize the same GPT.

Horizontal Externality: Improvements in GPT quality (z) benefit all application sectors.

"Too Late Innovation" (potentially): Each AS under-invests in its own complementary innovation (Ta) because it doesn't fully consider the positive impact its innovation has on other ASs (and thus, on GPT demand and future GPT quality).

Analogy to Public Goods: GPT quality (z) has some characteristics of a public good – non-rivalrous and non-excludable among application sectors.

Dynamics

- Use Markov Perfect Equilibrium Concept.
- Model the GPT producer and the applications as taking turns.
- Better forecasting / higher discount factor leads to higher technology levels.

Dynamics

- PC manufacturers knew about Intel's next-gen processors (e.g., Pentium).
- Knowledge allowed partial R&D before actual chip release.
- Information flow affected by institutional arrangements.
- Difficult technology forecasting leads to slower innovation
- Coordination capability impacts growth

Implications

- Importance of predictable demand. DOD? Government? FAANG?
- Importance of coordination. Notice some labs building applications in addition to the foundation models.
- AI Labs work with specific companies at application layer, e.g. OpenAI and Harvey.
- Importance of application layer. LLMs don’t increase innovation unless they are correctly plugged into production processes.
- Importance of capital, intermediate revenue for AI companies.

The task based model.

- Used in the work of Autor, Acemoglu, Restrepo, and others.
- Output:

$$Y = B(N)\left(\int_0^N y(z)^{\frac{\sigma-1}{\sigma}} dz\right)^{\frac{\sigma}{\sigma-1}}$$

- Each z is a task, and these models allow for increases in the number of tasks N.
- Tasks can be produced by labor or by capital (AI?).
- Acemoglu makes a bunch of simplifications to come up with a formula for how AI affects productivity (cost savings times share of tasks).

$$d \ln \mathrm{TFP} = \bar{\pi} \times \text{ GDP share of tasks impacted by AI.}$$

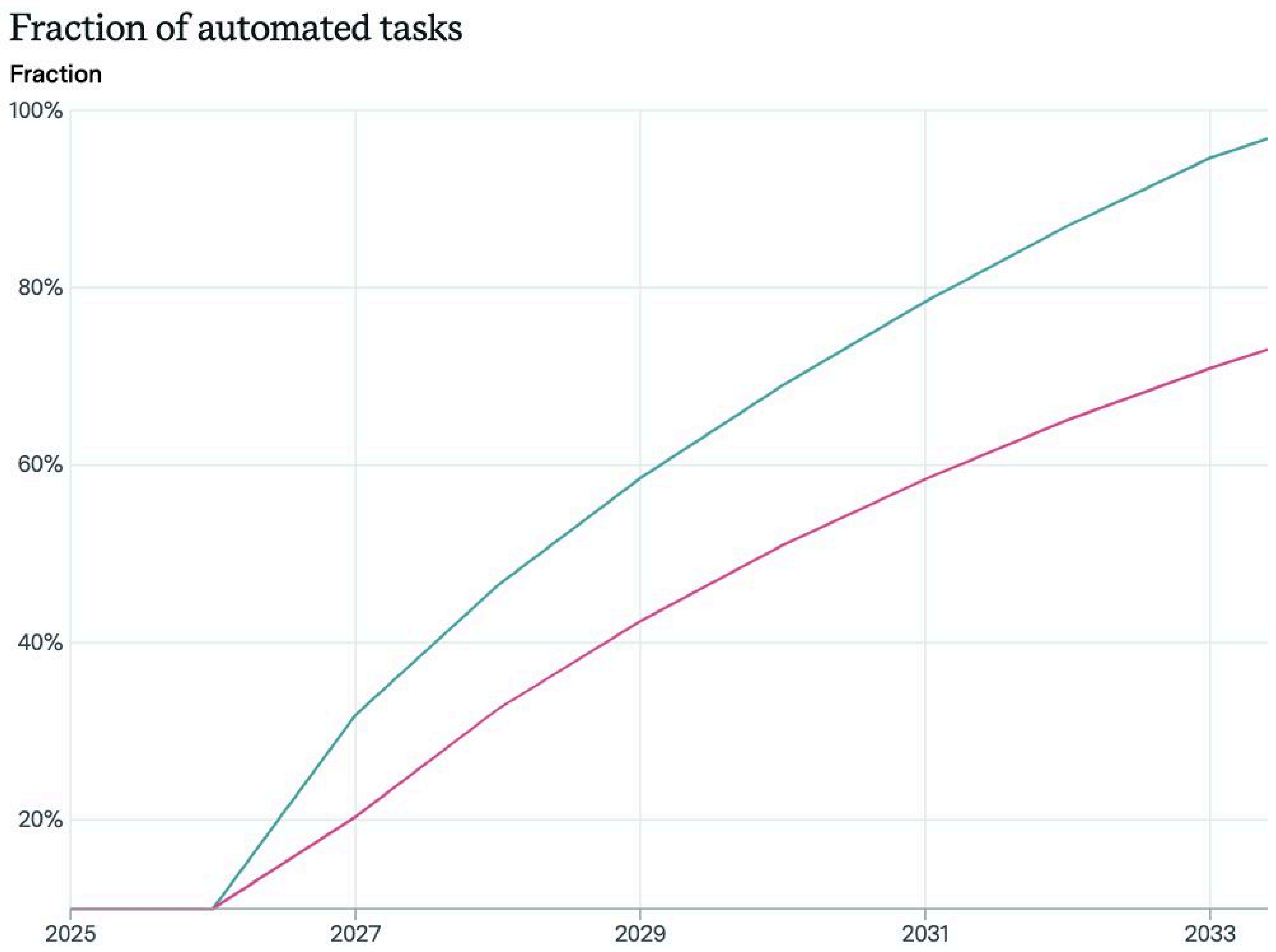
$$\begin{aligned} \text{TFP gains over the next 10 years} &= 0.046 \times 0.154 \\ &= 0.0071. \end{aligned}$$

Notice the disconnect

Acemoglu

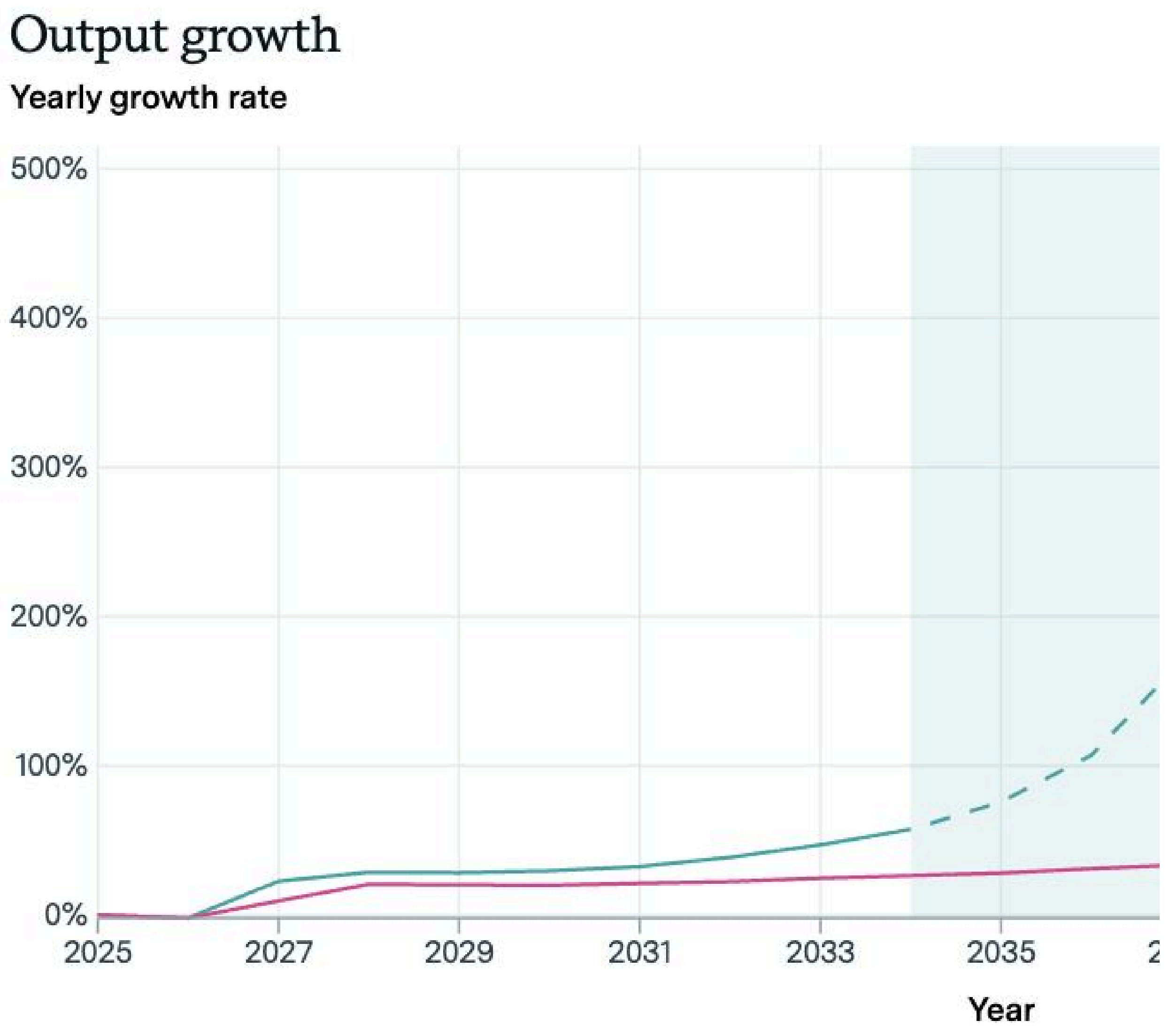
$$\text{TFP gains over the next 10 years} = 0.046 \times 0.154 = 0.0071.$$

EPOCH AI - GATE Model



20% growth in 2027

100% automation by 2035



Key question for all of us. My viewpoint.

- AI is going to affect every single part of knowledge work and eventually physical work.
- The biggest risks and opportunities for us as researchers:
  - Not using AI enough. It may be a better writer, presentation maker, coder, agenda setter, therapist, etc…
  - Doing research that is made obsolete by AI in < 5 years due to AI being able to do it or due to phenomena not existing any more. (E.g., narrow questions about ad copy design or about platforms whose business models will be destroyed).
  - Not investing correctly in skills, assets, etc…
- GDP growth implications of AI are likely to be backloaded. Coming up with and deploying new technologies will take time, especially since society is filled with frictions.
- Nonetheless, we will start seeing large productivity gains in at least some industries within 5 years. Think accounting, law, sales, programming, administrative work, already in customer support / translation.

Outline

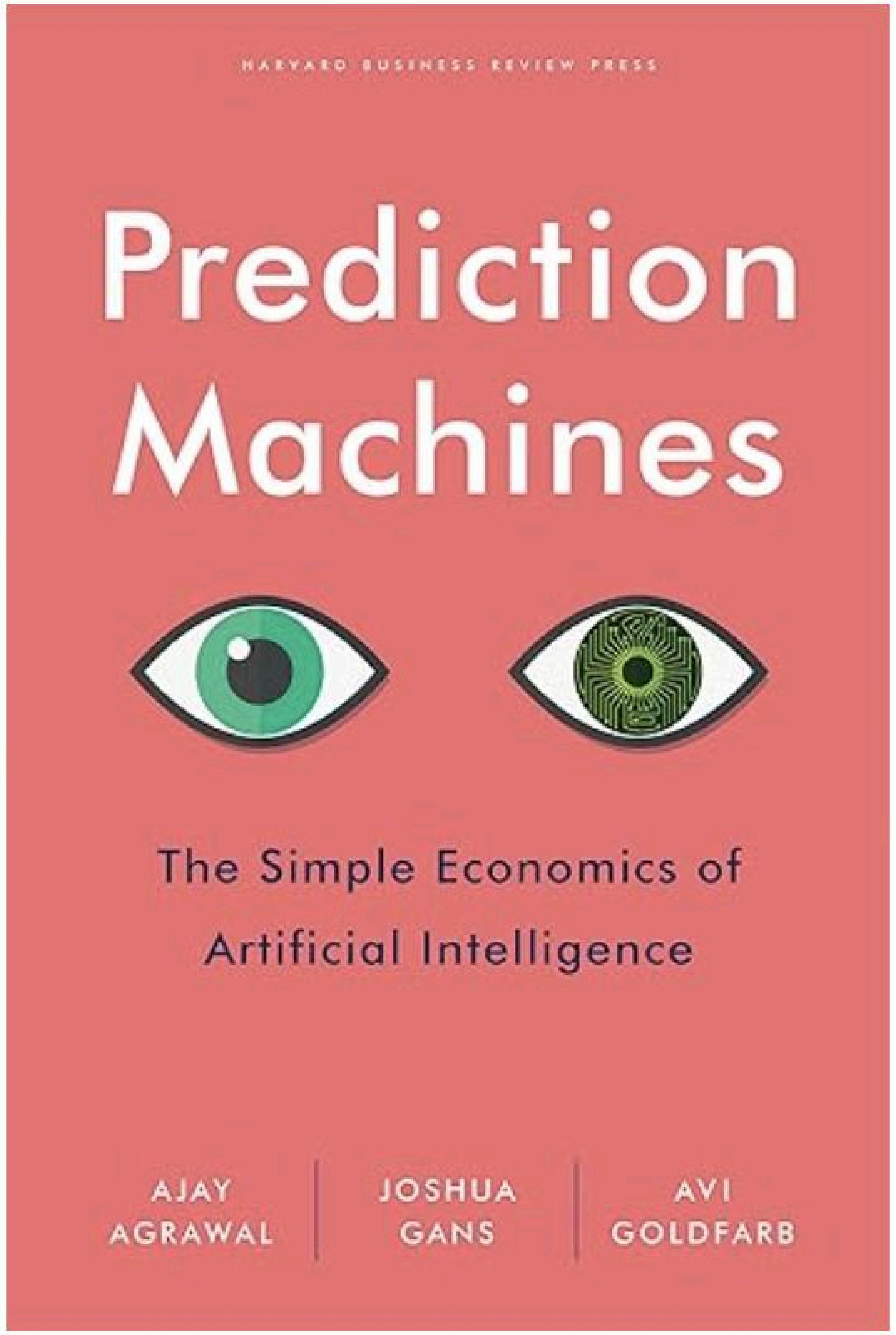
AGI - Situational Awareness

The Macroeconomics View

AI Narrow View - Prediction Policy Problems

Prediction Machines

- Economists view on AI circa a few years ago.
- AI is good at prediction but not good at judgment.
- Good to think about what parts of a decision problem are about prediction and what parts are about judgment.
- For example:
  - Is the research idea interesting enough to be publishable in a top journal? (prediction problem).
  - Do you work on the research idea given prediction? (judgment problem)



Kleinberg et al.

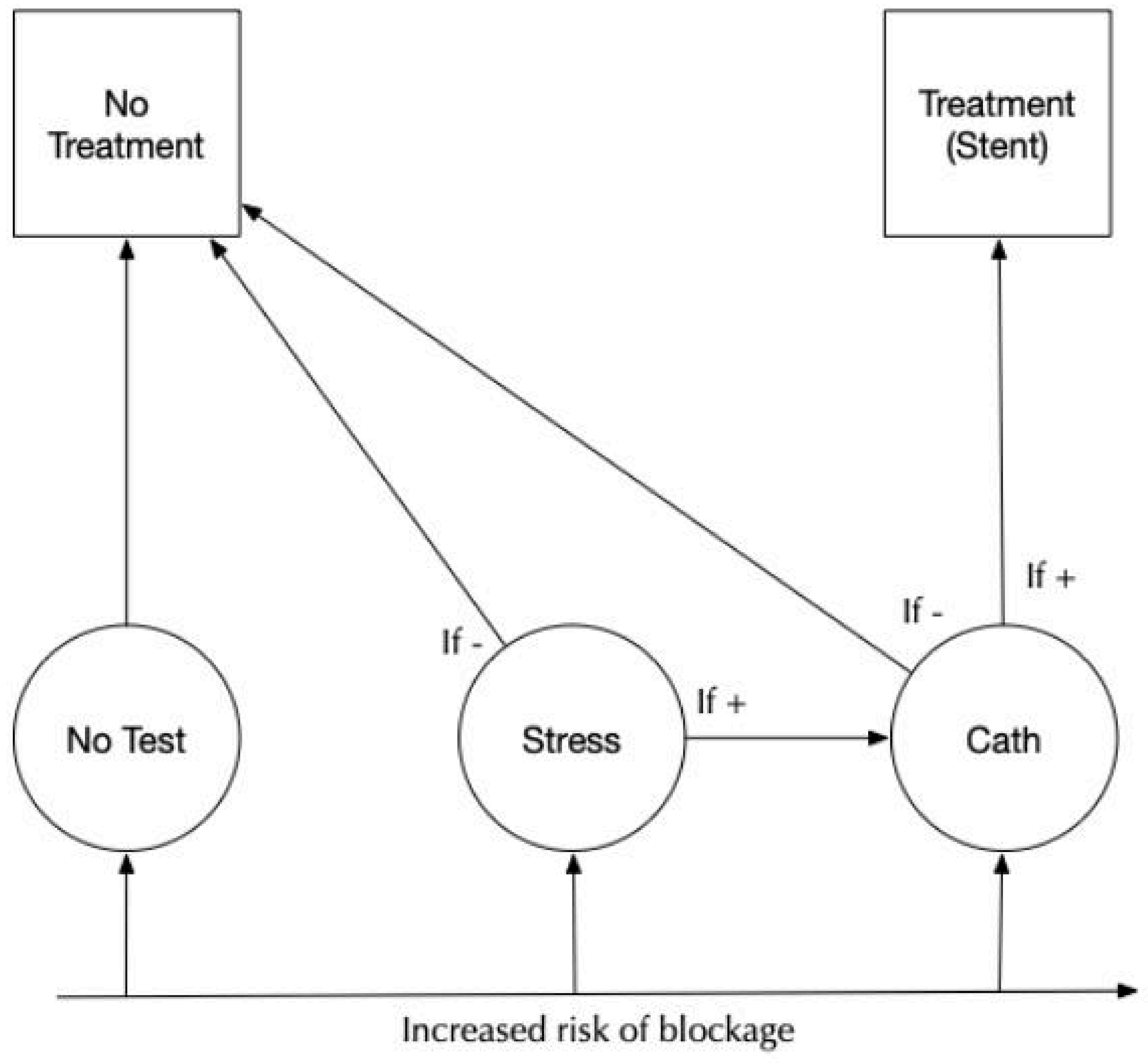
- Machine learning is good at predicting. For example, predicting the rain or whether someone will die soon.
- But machine learning isn't as good at causality. Does seeding clouds cause the rain? Does doing an operation reduce mortality.
- Most policy problems are a combination of both.

TABLE 1—RISKIEST JOINT REPLACEMENTS

Predicted mortality percentile	Observed mortality rate	Futile procedures averted	Futile spending ($ mill.)
1	0.435 (0.028)	1,984	30
2	0.422 (0.028)	3,844	58
5	0.358 (0.027)	8,061	121
10	0.242 (0.024)	10,512	158
20	0.152 (0.020)	12,317	185
30	0.136 (0.019)	16,151	242

Mullainathan and Obermeyer (2022, QJE)

- Full blown paper on using AI to improve productivity in healthcare.
- The setting is the emergency room (ER), where patients come in and need to be tested for a heart attack or not.
- Approach: Use ML to predict the likelihood of a positive test result and compare that to physician decisions.
- Use it to identify over and under testing.



Mullainathan and Obermeyer (2022, QJE)

- Why is this a non-trivial exercise?
  - Physicians observe factors not in the dataset.
  - Need to observe cost of not treating those who aren't tested. They may eventually have a heart attack or come back to the ER.
- Financial cost is important, since catheterization is a $30,000 procedure.
- Data: EHR records from a large academic hospital.

Table 1: Summary Statistics: Patient Characteristics

	All	Tested	Untested
N Patients	130,059	6,088	123,971
N Visits	246,874	7,320	239,554
Demographics			
Age, mean	42	58	42
	(0.033)	(0.146)	(0.033)
Female	0.611	0.459	0.616
	(<0.001)	(0.006)	(<0.001)
Black	0.262	0.216	0.264
	(<0.001)	(0.005)	(<0.001)
Hispanic	0.237	0.145	0.24
	(<0.001)	(0.004)	(<0.001)
White	0.436	0.588	0.432
	(<0.001)	(0.006)	(0.001)
Risk factors			
Past Heart Disease	0.121	0.391	0.113
	(<0.001)	(0.006)	(<0.001)
Diabetes	0.142	0.294	0.137
	(<0.001)	(0.005)	(<0.001)
Hypertension	0.251	0.513	0.243
	(<0.001)	(0.006)	(<0.001)
Cholesterol	0.162	0.417	0.155
	(<0.001)	(0.006)	(<0.001)
Any Risk Factor	0.36	0.625	0.351
	(<0.001)	(0.006)	(<0.001)
Triage Shifts			
Number of Shifts	5,925		
Patients per Shift	42		

Framework

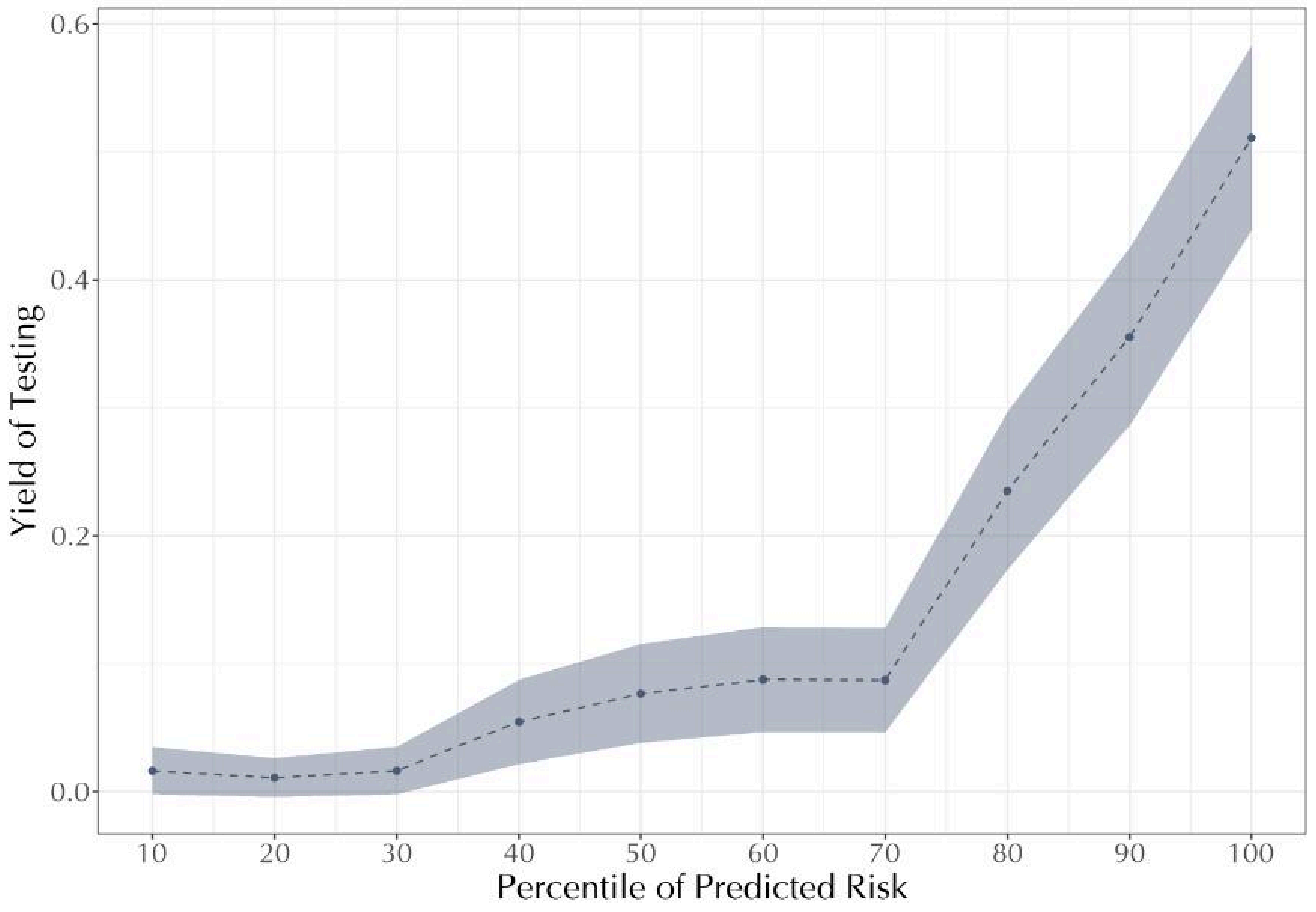
- Test if P(Blockage|X, Z) > cost
- Physicians estimate a probability of blockage h(X, Z) vs true P(B|X,Z).
- Z is private information.
- Mechanisms: physician error, moral hazard.
- How to get around not seeing Z? Use the time when a patient arrives as an exogenous shifter of likelihood of test.
    - Some shifts test more than others.

Table 1: Summary Statistics: Patient Characteristics

	All	Tested	Untested
N Patients	130,059	6,088	123,971
N Visits	246,874	7,320	239,554
Demographics			
Age, mean	42	58	42
	(0.033)	(0.146)	(0.033)
Female	0.611	0.459	0.616
	(<0.001)	(0.006)	(<0.001)
Black	0.262	0.216	0.264
	(<0.001)	(0.005)	(<0.001)
Hispanic	0.237	0.145	0.24
	(<0.001)	(0.004)	(<0.001)
White	0.436	0.588	0.432
	(<0.001)	(0.006)	(0.001)
Risk factors			
Past Heart Disease	0.121	0.391	0.113
	(<0.001)	(0.006)	(<0.001)
Diabetes	0.142	0.294	0.137
	(<0.001)	(0.005)	(<0.001)
Hypertension	0.251	0.513	0.243
	(<0.001)	(0.006)	(<0.001)
Cholesterol	0.162	0.417	0.155
	(<0.001)	(0.006)	(<0.001)
Any Risk Factor	0.36	0.625	0.351
	(<0.001)	(0.006)	(<0.001)
Triage Shifts			
Number of Shifts	5,925		
Patients per Shift	42		

What happens?

(a) Realized Yield of Testing



(b) Cost-Effectiveness of Testing

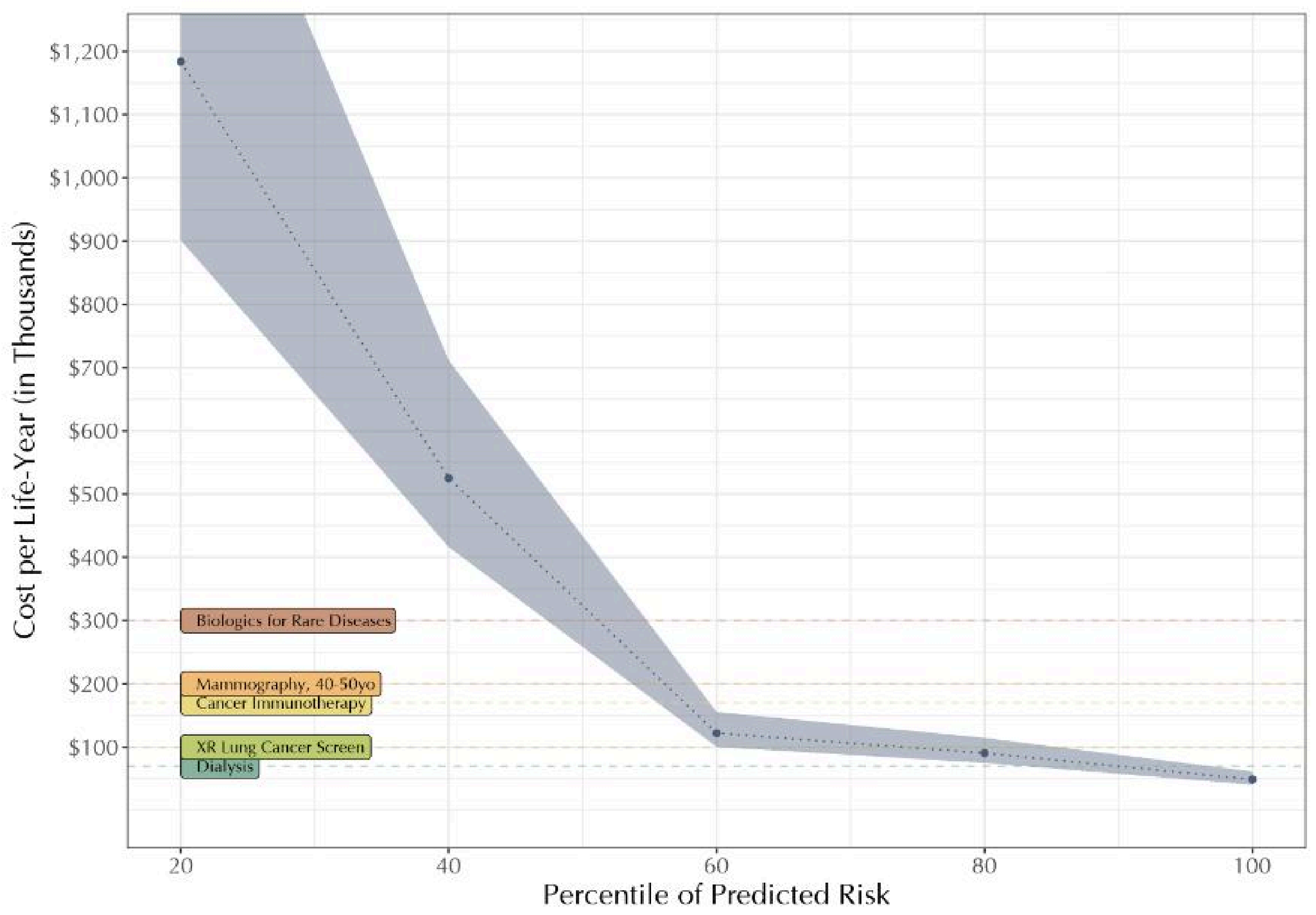
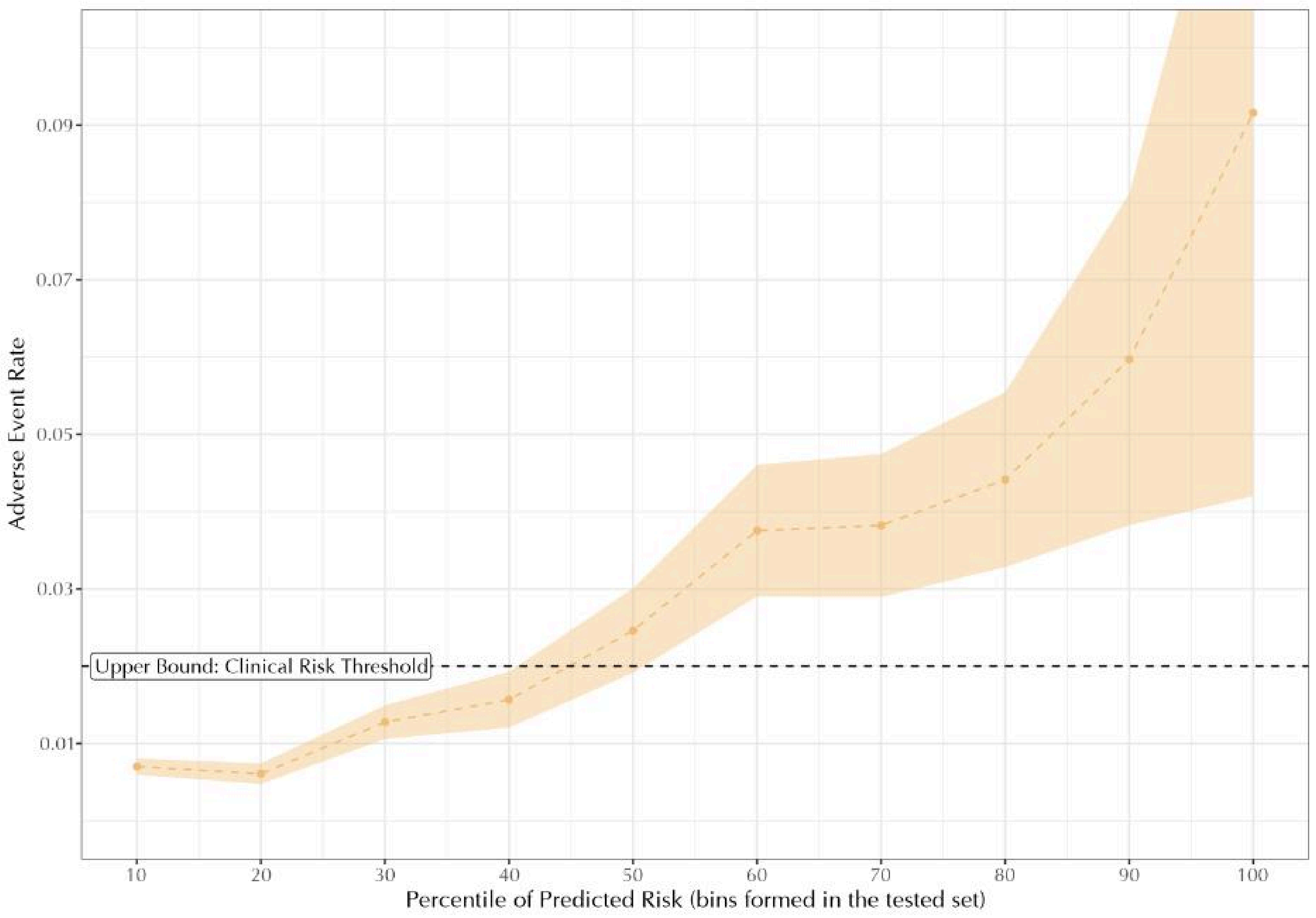


Figure 2: Adverse Events in Untested Patients (30 Days After Visits)

(a) Any Adverse Event



The triage identification strategy

Figure 4: Balance on Observables Across Triage Shfits

(a) Variation in Testing Rate and Observables, by Shift Testing Rate

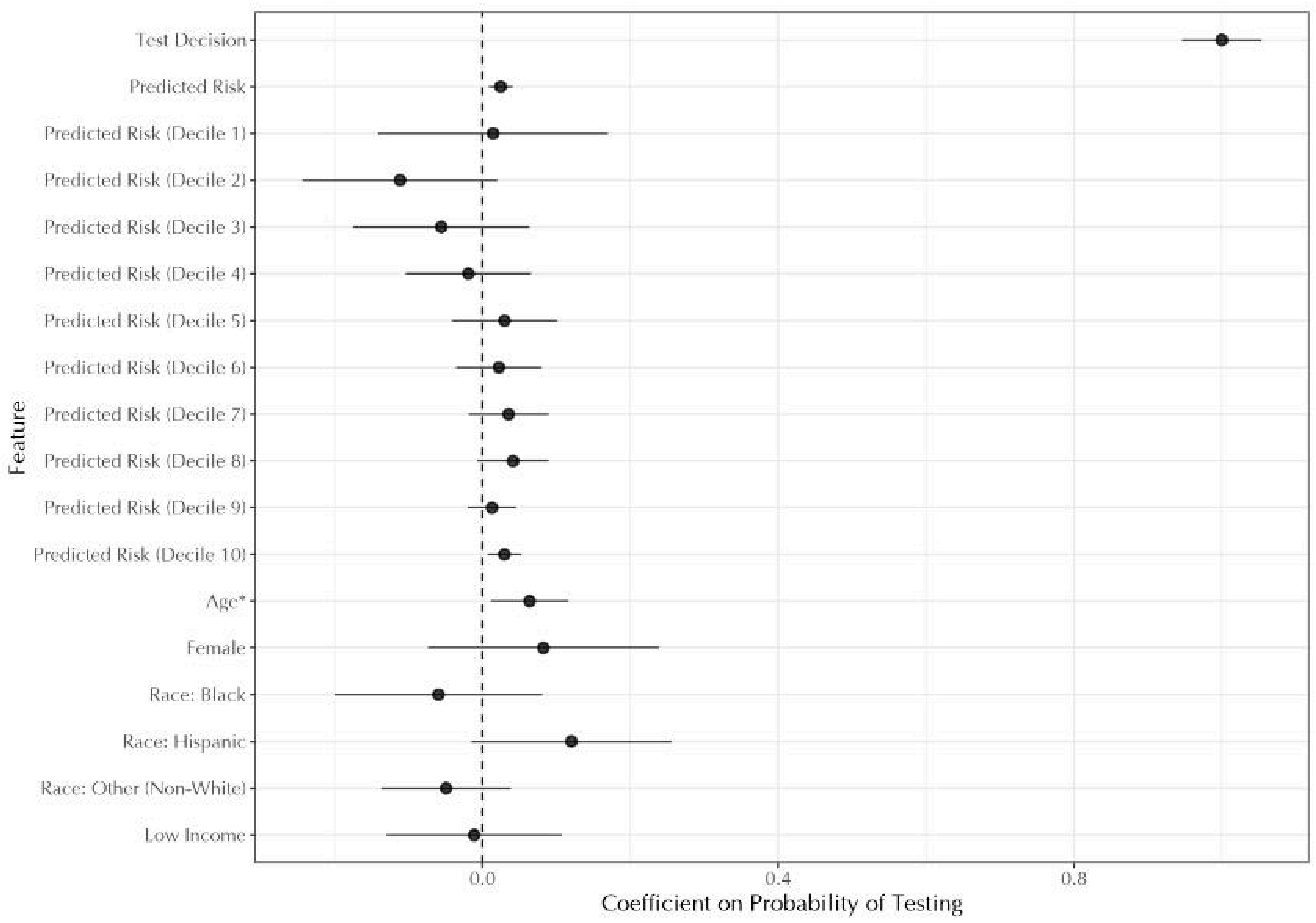


Table 5: Average Effect of Testing on Long-Term Adverse Events

	(1)	(2)	(3)	(4)
Testing Effect (Linear)	Adverse Event (31-365 days)	Diagnosed Event (31-365 days)	Death (31-365 days)	Death (365 days)
Shift Effect	−0.038	−0.007	−0.049**	−0.022
	(0.036)	(0.028)	(0.025)	(0.028)
Risk Control	Yes	Yes	Yes	Yes
Observations	213,484	213,484	213,484	213,484
R^2	0.010	0.003	0.012	0.021
Testing Effect (Quartiles)	Adverse Event (31-365 days)	Diagnosed Event (31-365 days)	Death (31-365 days)	Death (365 days)
Shift Q2	−0.040	0.015	−0.084	−0.086
	(0.100)	(0.079)	(0.069)	(0.078)
Shift Q3	0.140	0.160**	−0.021	−0.0001
	(0.100)	(0.079)	(0.069)	(0.078)
Shift Q4	−0.010	0.055	−0.116*	−0.068
	(0.100)	(0.080)	(0.069)	(0.078)
Risk Controls	Yes	Yes	Yes	Yes
Observations	213,484	213,484	213,484	213,484
R^2	0.006	0.001	0.008	0.015
Outcome Rates (%)	2.761	1.712	1.297	1.678

$^{*}p < .1, ^{**}p < .05, ^{***}p < .01$

Getting tested helps if you're high risk

Table 6: Effect of Testing on Long-Term Adverse Events By Predicted Risk

Risk Quintiles by Testing (Linear)	(1) Adverse Event (31-365 days)	(2) Diagnosed Event (31-365 days)	(3) Death (31-365 days)	(4) Death (0-365 days)
Testing	−0.037	−0.037	−0.028	−0.024
	(0.061)	(0.049)	(0.042)	(0.048)
Risk Q2 × Testing	0.070	0.083	0.010	0.032
	(0.084)	(0.066)	(0.058)	(0.065)
Risk Q3 × Testing	0.085	0.128	−0.019	0.011
	(0.102)	(0.081)	(0.070)	(0.080)
Risk Q4 × Testing	−0.316**	−0.129	−0.201*	−0.084
	(0.153)	(0.121)	(0.105)	(0.119)
Risk Q5 × Testing	−1.373***	−1.093***	−0.432**	−0.460**
	(0.275)	(0.219)	(0.190)	(0.215)
Risk Controls	Yes	Yes	Yes	Yes
R^2	0.010	0.003	0.012	0.021

Rest of the paper is behavioral economics + cost effectiveness

- Evidence of bounded rationality:
- Physicians use simpler models (k=49 variables) than optimal (k=224).
- Evidence of systematic biases:
  - Over-weight salient symptoms, especially chest pain
  - Over-weight representative symptoms (stereotypical of heart attack)
- Demographics biases (e.g., testing women more than warranted by risk)

“Putting this together with our estimate of over-testing above (49.1% of current tests), our counterfactual policy would cut testing on net by 11.8%—but of all the tests recommended under this policy, 42.3% would be high-value new tests, done for high-risk patients physicians are not currently testing.”

Next time: AI in the Wild

- Read Calvano et al.
- Read Lambrect & Tucker
- Intro to Agarwal et al. and Karlinsky-Shichor & Netzer (discussions).
- Assignment 4 is posted, due on April 23.