Leopold Aschenbrenner

SITUATIONAL AWARENESS

The Decade Ahead

Dedicated to Ilya Sutskever.

You can see the future first in San Francisco.

Over the past year, the talk of the town has shifted from $10 billion compute clusters to $100 billion clusters to trillion-dollar clusters. Every six months another zero is added to the boardroom plans. Behind the scenes, there's a fierce scramble to secure every power contract still available for the rest of the decade, every voltage transformer that can possibly be procured. American big business is gearing up to pour trillions of dollars into a long-unseen mobilization of American industrial might. By the end of the decade, American electricity production will have grown tens of percent; from the shale fields of Pennsylvania to the solar farms of Nevada, hundreds of millions of GPUs will hum.

The AGI race has begun. We are building machines that can think and reason. By 2025/26, these machines will outpace college graduates. By the end of the decade, they will be smarter than you or I; we will have superintelligence, in the true sense of the word. Along the way, national security forces not seen in half a century will be unleashed, and before long, The Project will be on. If we're lucky, we'll be in an all-out race with the CCP; if we're unlucky, an all-out war.

Everyone is now talking about AI, but few have the faintest glimmer of what is about to hit them. Nvidia analysts still think 2024 might be close to the peak. Mainstream pundits are stuck on the willful blindness of "it's just predicting the next word". They see only hype and business-as-usual; at most they entertain another internet-scale technological change.

Before long, the world will wake up. But right now, there are perhaps a few hundred people, most of them in San Francisco and the AI labs, that have *situational awareness*. Through whatever peculiar forces of fate, I have found myself amongst them. A few years ago, these people were derided as crazy—but they trusted the trendlines, which allowed them to correctly predict the AI advances of the past few years. Whether these people are also right about the next few years remains to be seen. But these are very smart people—the smartest people I have ever met—and they are the ones building this technology. Perhaps they will be an odd footnote in history, or perhaps they will go down in history like Szilard and Oppenheimer and Teller. If they are seeing the future even close to correctly, we are in for a wild ride.

Let me tell you what we see.

Contents

Reliably controlling AI systems much smarter than we are is an unsolved technical problem. And while it is a solvable problem, things could very easily go off the rails during a rapid intelligence explosion. Managing this will be extremely tense; failure could easily be catastrophic.

Superintelligence will give a decisive economic and military advantage. China isn't at all out of the game yet. In the race to AGI, the free world's very survival will be at stake. Can we maintain our preeminence over the authoritarian powers? And will we manage to avoid self-destruction along the way?

As the race to AGI intensifies, the national security state will get involved. The USG will wake from its slumber, and by 27/28 we'll get some form of government AGI project. No startup can handle superintelligence. Somewhere in a SCIF, the endgame will be on.

What if we're right?

I. From GPT-4 to AGI: Counting the OOMs

AGI by 2027 is strikingly plausible. <mark>GPT-2 to GPT-4 took us from ~preschooler to ~smart high-schooler abilities in 4 years</mark>. Tracing trendlines in compute (~0.5 orders of magnitude or OOMs/year), algorithmic efficiencies (~0.5 OOMs/year), and "unhobbling" gains (from chatbot to agent), <mark>we should expect another preschooler-to-high-schooler-sized qualitative jump by 2027.</mark>

> Look. The models, they just want to learn. You have to understand this. The models, they just want to learn.
>
> ILYA SUTSKEVER
> (circa 2015, via Dario Amodei)

GPT-4's CAPABILITIES came as a shock to many: an AI system that could write code and essays, could reason through difficult math problems, and ace college exams. A few years ago, most thought these were impenetrable walls.

But GPT-4 was merely the continuation of a decade of breakneck progress in deep learning. A decade earlier, models could barely identify simple images of cats and dogs; four years earlier, GPT-2 could barely string together semi-plausible sentences. Now we are rapidly saturating all the benchmarks we can come up with. And yet this dramatic progress has merely been the result of consistent trends in scaling up deep learning.

There have been people who have seen this for far longer. They were scoffed at, but all they did was trust the trendlines. The

trendlines are intense, and they were right. The models, they just want to learn; you scale them up, and they learn more.

I make the following claim: **it is strikingly plausible that by 2027, models will be able to do the work of an AI researcher/engineer**. That doesn't require believing in sci-fi; it just requires believing in straight lines on a graph.

Base Scaleup of Effective Compute

y-axis: Effective Compute (Normalized to GPT-4)

10^8, 10^7, 10^6, 10^5, 10^4, 10^3, 10^2, 10^1, 10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}

Automated AI Researcher/Engineer?

GPT-4: Smart High Schooler

GPT-3: Elementary Schooler

GPT-2: Preschooler

x-axis: 2018 2019 2020 2021 2022 2023 2024 2025 2026 2027 2028

Based on public estimates

SITUATIONAL AWARENESS | Leopold Aschenbrenner

In this piece, I will simply "count the OOMs" (OOM = order of magnitude, 10x = 1 order of magnitude): look at the trends in 1) *compute*, 2) *algorithmic efficiencies* (algorithmic progress that we can think of as growing "effective compute"), and 3) *"unhobbling" gains* (fixing obvious ways in which models are hobbled by default, unlocking latent capabilities and giving them tools, leading to step-changes in usefulness). We trace

Figure 1: Rough estimates of past and future scaleup of effective compute (both physical compute and algorithmic efficiencies), based on the public estimates discussed in this piece. As we scale models, they consistently get smarter, and by "counting the OOMs" we get a rough sense of what model intelligence we should expect in the (near) future. (This graph shows only the scaleup in base models; "unhobblings" are not pictured.)

the growth in each over four years before GPT-4, and what we should expect in the four years after, through the end of 2027. Given deep learning's consistent improvements for every OOM of effective compute, we can use this to project future progress.

Publicly, things have been quiet for a year since the GPT-4 release, as the next generation of models has been in the oven—leading some to proclaim stagnation and that deep learning is hitting a wall.[1] But by counting the OOMs, we get a peek at what we should actually expect.

The upshot is pretty simple. GPT-2 to GPT-4—from models that were impressive for sometimes managing to string together a few coherent sentences, to models that ace high-school exams—was not a one-time gain. We are racing through the OOMs extremely rapidly, and the numbers indicate we should expect another ~100,000x effective compute scaleup—resulting in another GPT-2-to-GPT-4-sized qualitative jump—over four years. Moreover, and critically, that doesn't just mean a better chatbot; picking the many obvious low-hanging fruit on "unhobbling" gains should take us from chatbots to agents, from a tool to something that looks more like drop-in remote worker replacements.

While the inference is simple, the implication is striking. Another jump like that very well could take us to AGI, to models as smart as PhDs or experts that can work beside us as coworkers. Perhaps most importantly, if these AI systems could automate AI research itself, that would set in motion intense feedback loops—the topic of the the next piece in the series.

Even now, barely anyone is pricing all this in. But situational awareness on AI isn't actually that hard, once you step back and look at the trends. If you keep being surprised by AI capabilities, just start counting the OOMs.

The last four years

WE HAVE MACHINES NOW that we can basically talk to like humans. It's a remarkable testament to the human capacity to adjust that this seems normal, that we've become inured to the pace of progress. But it's worth stepping back and looking at the progress of just the last few years.

GPT-2 to GPT-4

Let me remind you of how far we came in just the ~4 (!) years leading up to GPT-4.

GPT-2 (2019) ~ *preschooler*: "Wow, it can string together a few plausible sentences." A very-cherry-picked example of a semi-coherent story about unicorns in the Andes it generated was incredibly impressive at the time. And yet GPT-2 could barely count to 5 without getting tripped up;[2] when summarizing an article, it just barely outperformed selecting 3 random sentences from the article.[3]

[2] From SSC: "Janelle Shane asks GPT-2 its ten favorite animals:
   Prompt: My 10 favorite animals are: 1.
   My ten favorite animals are:
1. Zebras with a white scar on the back
2. Insiduous spiders and octopus
3. Frog with large leaves, hopefully black
4. Cockatiel with scales
5. Razorbill with wings hanging about 4 inches from one's face and a heart tattoo on a frog
3. Cockatric interlocking tetrabods that can be blind, cut, and eaten raw:
4. Black and white desert crocodiles living in sunlight
5. Zebra and many other pea bugs"

[3] From the GPT-2 paper, Section 3.6.

GPT-2 examples people found very impressive at the time

Task	**Reading Comprehension:** answer questions about given passages
Dataset	COQA
Examples	*The trophy doesn't fit into the brown suitcase because it is too large.*
	Correct answer: *it = trophy*
	Model answer: it = trophy
	The trophy doesn't fit into the brown suitcase because it is too small.
	Correct answer: *it = suitcase*
	Model answer: it = suitcase

System Prompt (human-written)   *For today's homework assignment, please describe the reasons for the US Civil War.*

Model Completion (machine-written, 10 tries)

```
By Donny Ferguson

It is easy to identify why the Civil War happened, because so
many people and so many books and so much television and films
tell us that it was the cause, that it has something to do with
race or economics or religion. There is much agreement that it
was essentially a war of slavery on behalf of capitalism, about a
century of slavery. But that's not what most people think of when
they think of the Civil War. Many people think of the war as
about states rights. You're not wrong about that, but it's kind
of misleading to say that the Civil War was a conflict between
states' rights and federalism. So let's try again. What's the
idea that has inspired so many people to go onto the battlefields
of Gettysburg and Antietam and Gettysburg and Petersburg and
Fredericksburg? The American idea of the republic--a notion of
limited government--is a great part of the history.
```

SITUATIONAL AWARENESS | Leopold Aschenbrenner

Comparing AI capabilities with human intelligence is difficult and flawed, but I think it's informative to consider the analogy

Figure 2: Some examples of what people found impressive about GPT-2 at the time. Left: GPT-2 does an ok job on extremely basic reading comprehension questions. Right: In a cherry-picked sample (best of 10 tries), GPT-2 can write a semi-coherent paragraph that says some semi-relevant things about the Civil War.

here, even if it's highly imperfect. GPT-2 was shocking for its command of language, and its ability to occasionally generate a semi-cohesive paragraph, or occasionally answer simple factual questions correctly. It's what would have been impressive for a preschooler.

GPT-3 (2020)[4] ~ *elementary schooler*: "Wow, with just some few-shot examples it can do some simple useful tasks." It started being cohesive over even multiple paragraphs much more consistently, and could correct grammar and do some very basic arithmetic. For the first time, it was also commercially useful in a few narrow ways: for example, GPT-3 could generate simple copy for SEO and marketing.

[4] I mean clunky old GPT-3 here, not the dramatically-improved GPT-3.5 you might know from ChatGPT.

GPT-3 examples people found very impressive at the time

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:
I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:
We screeghed at each other for several minutes and then we went outside and ate ice cream.

"The challenge is, you have to win a game of chess against me. If you win, then you may try to take the book. I will get the board." The door unlocks and opens.

> You go inside and play chess against the person.
You go inside and see a beautiful woman with long blond hair. She wears a purple dress and a silver tiara. You have seen this woman before. In fact, you have seen her many times before in pictures and statues.

> You try to remember who she is.
You have a sudden realization of who this person is. This is the lost queen of the kingdom! She was thought to have been killed by bandits, along with her husband, but it looks like she survived.

> You express your surprise and ask her why she let people believe she is dead.
"I am surprised you did not know, with all your wizard knowledge. Let's just say I am hiding from someone. If people thought I was dead, then they wouldn't be looking for me." She says.

> You ask her who she is hiding from.
"If you want to challenge me to a game, then play. If you want to pry into my personal life, then you can leave."
You decide not to press the issue, so you sit down and start a game of chess.

Describe a layout.

Just describe any layout you want, and it'll try to render below!

a button that looks like a watermelon

Generate

`<button style={{backgroundColor: 'pink', border: '2px solid green', borderRadius: '50%', padding: 20, width: 100, height: 100}}>Watermelon</button>`

Watermelon

SITUATIONAL AWARENESS | Leopold Aschenbrenner

Again, the comparison is imperfect, but what impressed people about GPT-3 is perhaps what would have been impressive for an elementary schooler: it wrote some basic poetry, could tell richer and coherent stories, could start to do rudimentary

Figure 3: Some examples of what people found impressive about GPT-3 at the time. Top: After a simple instruction, GPT-3 can use a made-up word in a new sentence. Bottom-left: GPT-3 can engage in rich storytelling back-and-forth. Bottom-right: GPT-3 can generate some very simple code.

coding, could fairly reliably learn from simple instructions and demonstrations, and so on.

GPT-4 (2023) ~ *smart high schooler*: "Wow, it can write pretty sophisticated code and iteratively debug, it can write intelligently and sophisticatedly about complicated subjects, it can reason through difficult high-school competition math, it's beating the vast majority of high schoolers on whatever tests we can give it, etc." From code to math to Fermi estimates, it can think and reason. GPT-4 is now useful in my daily tasks, from helping write code to revising drafts.

GPT-4 examples people found very impressive at the time

Figure 4: Some of what people found impressive about GPT-4 when it was released, from the "Sparks of AGI" paper. Top: It's writing very complicated code (producing the plots shown in the middle) and can reason through nontrivial math problems. Bottom-left: Solving an AP math problem. Bottom-right: Solving a fairly complex coding problem. More interesting excerpts from that exploration of GPT-4's capabilities here.

On everything from AP exams to the SAT, GPT-4 scores better than the vast majority of high schoolers.

Of course, even GPT-4 is still somewhat uneven; for some tasks it's much better than smart high-schoolers, while there are other tasks it can't yet do. That said, I tend to think most of these limitations come down to obvious ways models are still hobbled, as I'll discuss in-depth later. The raw intelligence is (mostly) there, even if the models are still artificially constrained; it'll take extra work to unlock models being able to fully apply that raw intelligence across applications.

GPT-2   GPT-3   GPT-4

Preschooler   Elementary   Smart High   ???
               Schooler   Schooler

SITUATIONAL AWARENESS | Leopold Aschenbrenner
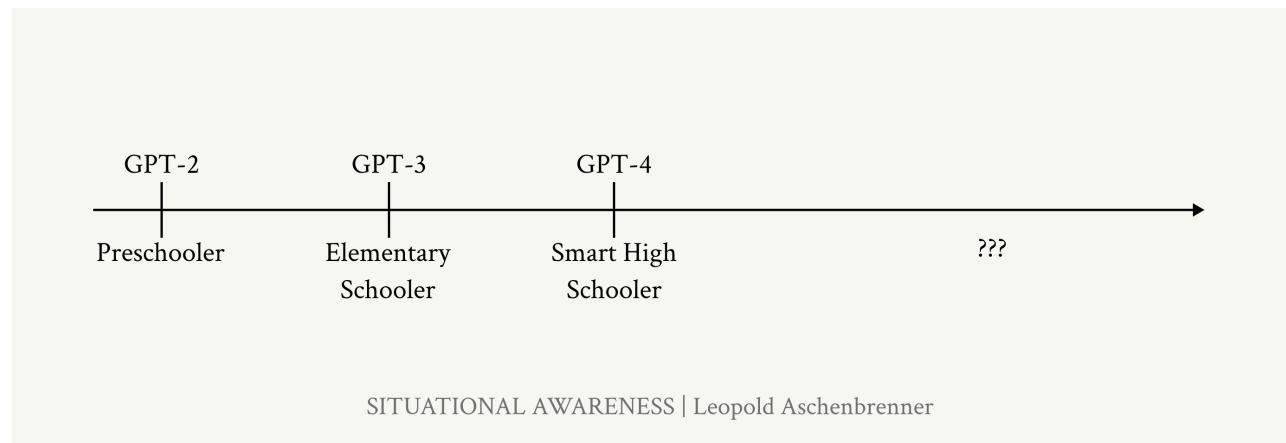
Figure 5: Progress over just four years. Where are **you** on this line?

The trends in deep learning

The pace of deep learning progress in the last decade has simply been extraordinary. A mere decade ago it was revolutionary for a deep learning system to identify simple images. Today, we keep trying to come up with novel, ever harder tests, and yet each new benchmark is quickly cracked. It used to take decades to crack widely-used benchmarks; now it feels like mere months.

We're literally running out of benchmarks. As an anecdote, my friends Dan and Collin made a benchmark called MMLU a few years ago, in 2020. They hoped to finally make a benchmark that would stand the test of time, equivalent to all the hardest exams we give high school and college students. Just three years later, it's basically solved: models like GPT-4 and Gemini

Test scores of AI systems on various capabilities relative to human performance

Within each domain, the initial performance of the AI is set to −100. Human performance is used as a baseline, set to zero. When the AI's performance crosses the zero line, it scored more points than humans.
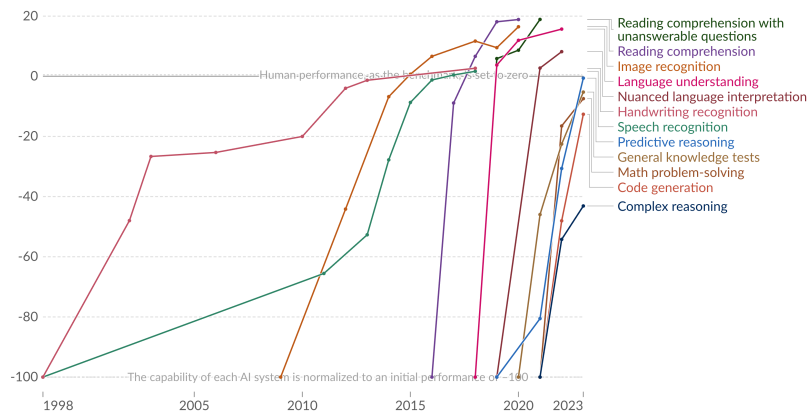
Data source: Kiela et al. (2023) OurWorldInData.org/artificial-intelligence | CC BY
Note: For each capability, the first year always shows a baseline of −100, even if better performance was recorded later that year.

Figure 6: Deep learning systems are rapidly reaching or exceeding human-level in many domains. Graphic: Our World in Data

get ~90%.

More broadly, GPT-4 mostly cracks all the standard high school and college aptitude tests (Figure 7).[5]

Or consider the MATH benchmark, a set of difficult mathematics problems from high-school math competitions.[6] When the benchmark was released in 2021, GPT-3 only got ~5% of problems right. And the original paper noted: "Moreover, we find that simply increasing budgets and model parameter counts will be impractical for achieving strong mathematical reasoning if scaling trends continue [...]. To have more traction on mathematical problem solving we will likely need new algorithmic advancements from the broader research community"—we would need fundamental new breakthroughs to solve MATH, or so they thought. A survey of ML researchers predicted minimal progress over the coming years (Figure 8);[7] and yet within just a year (by mid-2022), the best models went from ~5% to 50% accuracy; now, MATH is basically solved, with recent performance over 90%.

[5] And no, these tests aren't in the training set. AI labs put real effort into ensuring these evals are uncontaminated, because they need good measurements in order to do good science. A recent analysis on this by ScaleAI confirmed that the leading labs aren't overfitting to the benchmarks (though some smaller LLM developers might be juicing their numbers).

[6] In the original paper, it was noted: "We also evaluated humans on MATH, and found that a computer science PhD student who does not especially like mathematics attained approximately 40% on MATH, while a three-time IMO gold medalist attained 90%, indicating that MATH can be challenging for humans as well."

[7] A coauthor notes: "When our group first released the MATH dataset, at least one [ML researcher colleague] told us that it was a pointless dataset because it was too far outside the range of what ML models could accomplish (indeed, I was somewhat worried about this myself)."

Performance on common exams
(percentile compared to human test-takers)

	GPT-4 (2023)	GPT-3.5 (2022)
Uniform Bar Exam	90th	10th
LSAT	88th	40th
SAT	97th	87th
GRE (Verbal)	99th	63rd
GRE (Quantitative)	80th	25th
US Biology Olympiad	99th	32nd
AP Calculus BC	51st	3rd
AP Chemistry	80th	34th
AP Macroeconomics	92nd	40th
AP Statistics	92nd	51st

SITUATIONAL AWARENESS | Leopold Aschenbrenner

Figure 7: GPT-4 scores on standardized tests. Note also the large jump from GPT-3.5 to GPT-4 in human percentile on these tests, often from well below the median human to the very top of the human range. (And this is GPT-3.5, a fairly recent model released less than a year before GPT-4, not the clunky old elementary-school-level GPT-3 we were talking about earlier!)
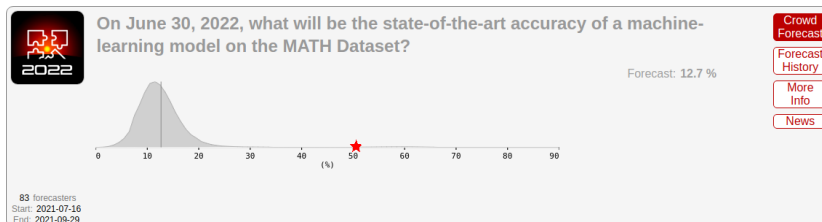
Figure 8: Gray: Professional forecasts, made in August 2021, for June 2022 performance on the MATH benchmark (difficult mathematics problems from high-school math competitions). Red star: actual state-of-the-art performance by June 2022, far exceeding even the upper range forecasters gave. The median ML researcher was even more pessimistic.

Over and over again, year after year, skeptics have claimed "deep learning won't be able to do X" and have been quickly proven wrong.[8] *If there's one lesson we've learned from the past decade of AI, it's that you should never bet against deep learning*.

Now the hardest unsolved benchmarks are tests like GPQA, a set of PhD-level biology, chemistry, and physics questions. Many of the questions read like gibberish to me, and even PhDs in other scientific fields spending 30+ minutes with Google barely score above random chance. Claude 3 Opus currently gets ~60%,[9] compared to in-domain PhDs who get ~80%—and I expect this benchmark to fall as well, in the next generation or two.

[8] Here's Yann LeCun predicting in 2022 that even GPT-5000 won't be able to reason about physical interactions with the real world; GPT-4 obviously does it with ease a year later.

Here's Gary Marcus's walls predicted after GPT-2 being solved by GPT-3, and the walls he predicted after GPT-3 being solved by GPT-4.

Here's Prof. Bryan Caplan losing his first-ever public bet (after previously famously having a perfect public betting track record). In January 2023, after GPT-3.5 got a D on his economics midterm, Prof. Caplan bet Matthew Barnett that no AI would get an A on his economics midterms by 2029. Just two months later, when GPT-4 came out, it promptly scored an A on his midterm (and it would have been one of the highest scores in his class).

[9] On the diamond set, majority voting of the model trying 32 times with chain-of-thought.

Chemistry (general)

A reaction of a liquid organic compound, which molecules consist of carbon and hydrogen atoms, is performed at 80 centigrade and 20 bar for 24 hours. In the proton nuclear magnetic resonance spectrum, the signals with the highest chemical shift of the reactant are replaced by a signal of the product that is observed about three to four units downfield. Compounds from which position in the periodic system of the elements, which are also used in the corresponding large-scale industrial process, have been mostly likely initially added in small amounts?
A) A metal compound from the fifth period.
B) A metal compound from the fifth period and a non-metal compound from the third period.
C) A metal compound from the fourth period.
D) A metal compound from the fourth period and a non-metal compound from the second period.

Organic Chemistry

Methylcyclopentadiene (which exists as a fluxional mixture of isomers) was allowed to react with methyl isoamyl ketone and a catalytic amount of pyrrolidine. A bright yellow, cross-conjugated polyalkenyl hydrocarbon product formed (as a mixture of isomers), with water as a side product. These products are derivatives of fulvene. This product was then allowed to react with ethyl acrylate in a 1:1 ratio. Upon completion of the reaction, the bright yellow color had disappeared. How many chemically distinct isomers make up the final product (not counting stereoisomers)?
A) 2
B) 16
C) 8
D) 4

Genetics

If a sperm from species A is injected into an egg from species B and both species have the same number of chromosomes, what would be the main cause of the resulting zygote mortality?
A) Species specific zona pellucida proteins on the egg cannot bind sperms from a different species.
B) Epistatic interactions between the genes of different species
C) Chromosomal incompatibilities will cause failure of meiosis leading to death of zygote.
D) Chromosomal recombination will not occur in different species.

Molecular Biology

A scientist studies the stress response of barley to increased temperatures and finds a protein which contributes to heat tolerance through the stabilisation of cell membrane. The scientist is very happy and wants to create a heat-tolerant cultivar of diploid wheat. Using databases, they find a heat tolerance protein homologue and start analysing its accumulation under heat stress. Soon enough, the scientist discovers this protein is not synthesised in the wheat cultivar they study. There are many possible reasons for such behaviour, including:
A) A miRNA targets the protein, which makes exonucleases cut it immediately after the end of translation and before processing in ER
B) Trimethylation of lysine of H3 histone in position 27 at the promoter of the gene encoding the target protein
C) A stop-codon occurs in the 5'-UTR region of the gene encoding the target protein
D) The proteolysis process disrupts a quaternary structure of the protein, preserving only a tertiary structure

---

Astrophysics

Astronomers are studying a star with a Teff of approximately 6000 K. They are interested in spectroscopically determining the surface gravity of the star using spectral lines (EW < 100 mA) of two chemical elements, El1 and El2. Given the atmospheric temperature of the star, El1 is mostly in the neutral phase, while El2 is mostly ionized. Which lines are the most sensitive to surface gravity for the astronomers to consider?
A) El2 I (neutral)
B) El1 II (singly ionized)
C) El2 II (singly ionized)
D) El1 I (neutral)

Quantum Mechanics

Suppose we have a depolarizing channel operation given by $E(\rho)$. The probability, p, of the depolarization state represents the strength of the noise. If the Kraus operators of the given state are $A_0 = \sqrt{1 - \frac{3p}{4}}$, $A_1 = \sqrt{\frac{p}{4}}X$, $A_2 = \sqrt{\frac{p}{4}}Y$, and $A_3 = \sqrt{\frac{p}{4}}Z$. What could be the correct Kraus Representation of the state $E(\rho)$?
A) $E(\rho) = (1 - p)\rho + \frac{p}{3}X\rho X + \frac{p}{3}Y\rho Y + \frac{p}{3}Z\rho Z$
B) $E(\rho) = (1 - p)\rho + \frac{p}{3}X\rho^2 X + \frac{p}{3}Y\rho^2 Y + \frac{p}{3}Z\rho^2 Z$
C) $E(\rho) = (1 - p)\rho + \frac{p}{4}X\rho X + \frac{p}{4}Y\rho Y + \frac{p}{4}Z\rho Z$
D) $E(\rho) = (1 - p)\rho^2 + \frac{p}{3}X\rho^2 X + \frac{p}{3}Y\rho^2 Y + \frac{p}{3}Z\rho^2 Z$

Table 1: Six example questions from the dataset, two each from subdomains of chemistry, biology, and physics (respectively).

Figure 9: Example GPQA questions. Models are already better at this than I am, and we'll probably crack expert-PhD-level soon. . .

Counting the OOMs

How did this happen? The magic of deep learning is that it *just works*—and the trendlines have been astonishingly consistent, despite naysayers at every turn.

Base compute   4x compute   32x compute
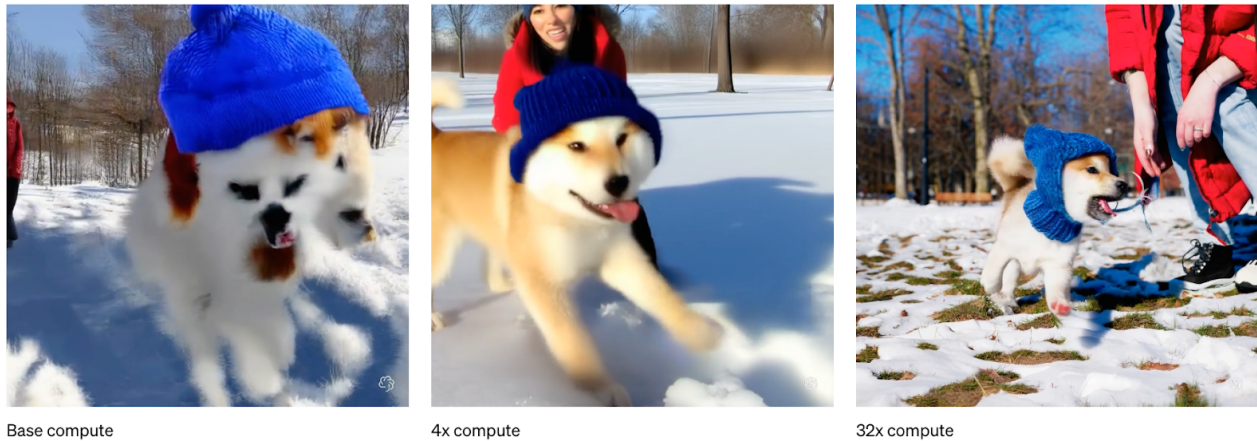
With each OOM of effective compute, models predictably, reliably get better.[10] If we can count the OOMs, we can (roughly, qualitatively) extrapolate capability improvements.[11] That's how a few prescient individuals saw GPT-4 coming.

We can decompose the progress in the four years from GPT-2 to GPT-4 into three categories of scaleups:

1. **COMPUTE**: We're using much bigger computers to train these models.

2. **ALGORITHMIC EFFICIENCIES**: There's a continuous trend of algorithmic progress. Many of these act as "compute multipliers," and we can put them on a unified scale of growing *effective compute*.

3. **"UNHOBBLING" GAINS**: By default, models learn a lot of amazing raw capabilities, but they are hobbled in all sorts of dumb ways, limiting their practical value. With simple algorithmic improvements like reinforcement learning from human feedback (RLHF), chain-of-thought (CoT), tools, and scaffolding, we can unlock significant latent capabilities.

Figure 10: The effects of scaling compute, in the example of OpenAI Sora.

[10] And it's worth noting *just how consistent* these trendlines are. Combining the original scaling laws paper with some of the estimates on compute and compute efficiency scaling since then implies a consistent scaling trend for *over 15 orders of magnitude* (over 1,000,000,000,000,000x in effective compute)!

[11] A common misconception is that scaling only holds for perplexity loss, but we see very clear and consistent scaling behavior on downstream performance on benchmarks as well. It's usually just a matter of finding the right log-log graph. For example, in the GPT-4 blog post, they show consistent scaling behavior for performance on coding problems over 6 OOMs (1,000,000x) of compute, using MLPR (mean log pass rate).

The "Are Emergent Abilities a Mirage?" paper makes a similar point; with the right choice of metric, there is almost always a consistent trend for performance on downstream tasks.

More generally, the "scaling hypothesis" qualitative observation—very clear trends on model capability with scale—predates loss-scaling-curves; the "scaling laws" work was just a formal measurement of this.

We can "count the OOMs" of improvement along these axes: that is, trace the scaleup for each in units of effective compute. 3x is 0.5 OOMs; 10x is 1 OOM; 30x is 1.5 OOMs; 100x is 2 OOMs; and so on. We can also look at what we should expect on top of GPT-4, from 2023 to 2027.

I'll go through each one-by-one, but the upshot is clear: we are rapidly racing through the OOMs. There are potential headwinds in the data wall, which I'll address—but overall, it seems likely that we should expect another GPT-2-to-GPT-4-sized jump, on top of GPT-4, by 2027.

Compute

I'll start with the most commonly-discussed driver of recent progress: throwing (a lot) more compute at models.

Many people assume that this is simply due to Moore's Law. But even in the old days when Moore's Law was in its heyday, it was comparatively *glacial*—perhaps 1-1.5 OOMs per decade. We are seeing much more rapid scaleups in compute—close to 5x the speed of Moore's law—instead because of mammoth *investment*. (Spending even a million dollars on a single model used to be an outrageous thought nobody would entertain, and now that's pocket change!)

Table 1: Estimates of compute for GPT-2 to GPT-4 by Epoch AI.

Model	Estimated Compute	Growth
GPT-2 (2019)	~4e21 FLOP	
GPT-3 (2020)	~3e23 FLOP	+ ~2 OOMs
GPT-4 (2023)	8e24 to 4e25 FLOP	+ ~1.5–2 OOMs

We can use public estimates from Epoch AI (a source widely respected for its excellent analysis of AI trends) to trace the compute scaleup from 2019 to 2023. GPT-2 to GPT-3 was a quick scaleup; there was a large overhang of compute, scaling from a smaller experiment to using an entire datacenter to train a large language model. With the scaleup from GPT-3 to GPT-4, we transitioned to the modern regime: having to build an entirely new (much bigger) cluster for the next model. And yet

the dramatic growth continued. Overall, Epoch AI estimates suggest that GPT-4 training used ~3,000x-10,000x more raw compute than GPT-2.

In broad strokes, this is just the continuation of a longer-running trend. For the last decade and a half, primarily because of broad scaleups in investment (and specializing chips for AI workloads in the form of GPUs and TPUs), the training compute used for frontier AI systems has grown at roughly ~0.5 OOMs/year.

Training compute of notable models
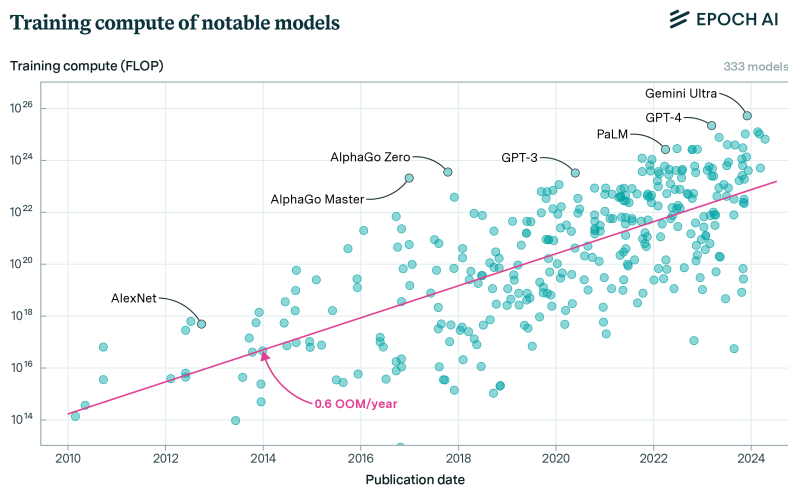
Figure 11: Training compute of notable deep learning models over time. Source: Epoch AI.

The compute scaleup from GPT-2 to GPT-3 in a year was an unusual overhang, but all the indications are that the longer-run trend will continue. The SF-rumor-mill is abreast with dramatic tales of huge GPU orders. The investments involved will be extraordinary—but they are in motion. I go into this more later in the series, in IIIa. Racing to the Trillion-Dollar Cluster; based on that analysis, an additional 2 OOMs of compute (a cluster in the $10s of billions) seems very likely to happen by the end of 2027; even a cluster closer to +3 OOMs of compute ($100 billion+) seems plausible (and is rumored to be in the works at Microsoft/OpenAI).

Algorithmic efficiencies

While massive investments in compute get all the attention, algorithmic progress is probably a similarly important driver of progress (and has been dramatically underrated).

To see just how big of a deal algorithmic progress can be, consider the following illustration (Figure 12) of the drop in price to attain ~50% accuracy on the MATH benchmark (high school competition math) over just *two years.* (For comparison, a computer science PhD student who didn't particularly like math scored 40%, so this is already quite good.) ==Inference efficiency improved by nearly 3 OOMs—1,000x—in less than two years.==

Relative (inference) cost of ~50% performance on the MATH benchmark

SITUATIONAL AWARENESS | Leopold Aschenbrenner

Figure 12: Rough estimate on relative inference cost of attaining ~50% MATH performance.

Calculations below.
Gemini 1.5 Flash scores 54.9% on MATH, and costs $0.35/$1.05 (input/output) per million tokens. GPT-4 scored 42.5% on MATH prelease and 52.9% on MATH in early 2023, and cost $30/$60 (input/output) per million tokens; that's 85x/57x (input/output) more expensive per token than Gemini 1.5 Flash. To be conservative, I use an estimate of 30x cost decrease above (accounting for Gemini 1.5 Flash possibly using more tokens to reason through problems).
Minerva540B scores 50.3% on MATH, using majority voting among 64 samples. A knowledgeable friend estimates the base model here is probably 2-3x more expensive to inference than GPT-4. However, Minerva seems to use somewhat fewer tokens per answer on a quick spot check. More importantly, Minerva needed 64 samples to achieve that performance, naively implying a 64x multiple on cost if you e.g. naively ran this via an inference API. In practice, prompt tokens can be cached when running an eval; given a few-shot prompt, prompt tokens are likely a majority of the cost, even accounting for output tokens. Supposing output tokens are a third of the cost for getting a single sample, that would imply only a ~20x increase in cost from the maj@64 with caching. To be conservative, I use the rough number of a 20x cost decrease in the above (even if the naive decrease in inference cost from running this via an API would be larger).

Though these are numbers just for inference efficiency (which may or may not correspond to training efficiency improvements, where numbers are harder to infer from public data), they make clear there is an *enormous* amount of algorithmic progress possible and happening.

In this piece, I'll separate out two kinds of algorithmic progress. Here, I'll start by covering "within-paradigm" algorithmic improvements—those that simply result in better base models, and that straightforwardly act as *compute efficiencies* or *compute multipliers*. For example, a better algorithm might allow us to achieve the same performance but with 10x less training compute. In turn, that would act as a 10x (1 OOM) increase in *effective compute*. (Later, I'll cover "unhobbling," which you can think of as "paradigm-expanding/application-expanding" algorithmic progress that unlocks capabilities of base models.)

If we step back and look at the long-term trends, we seem to find new algorithmic improvements at a fairly consistent rate. Individual discoveries seem random, and at every turn, there seem insurmountable obstacles—but the long-run trendline is predictable, a straight line on a graph. Trust the trendline.

We have the best data for ImageNet (where algorithmic research has been mostly public and we have data stretching back a decade), for which we have consistently improved compute efficiency by roughly ~0.5 OOMs/year across the 9-year period between 2012 and 2021.

(a) Pareto frontiers in data and compute for AlexNet performance

(b) Pareto frontiers in data and compute for ResNeXt-101 performance

(c) Pareto frontiers in data and compute for ViT-e performance
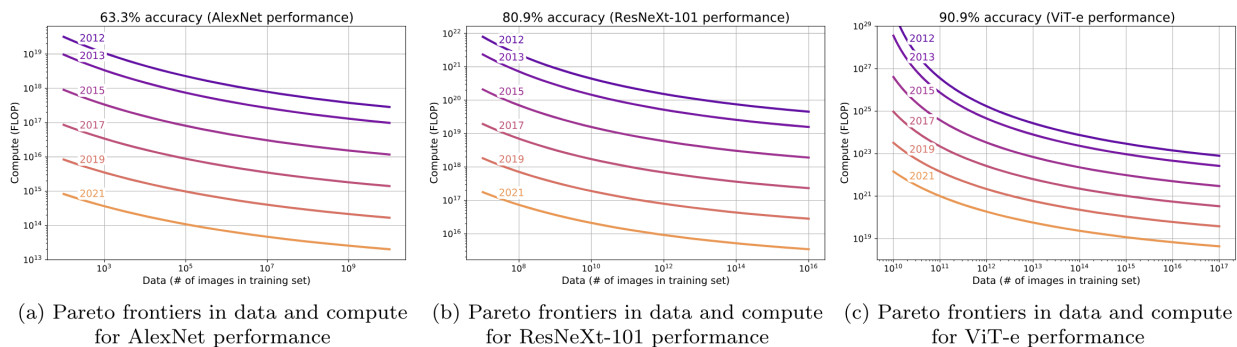
Figure 1. Pareto frontiers for training models to achieve performance of well-known models over time.

That's a huge deal: that means 4 years later, we can achieve the same level of performance for ~100x less compute (and concomitantly, much higher performance for the same compute!).

Unfortunately, since labs don't publish internal data on this, it's harder to measure algorithmic progress for frontier LLMs

Figure 13: We can measure algorithmic progress: how much less compute is needed in 2021 compared to 2012 to train a model with the same performance? We see a trend of ~0.5 OOMs/year of algorithmic efficiency. Source: Erdil and Besiroglu 2022.

over the last four years. EpochAI has new work replicating their results on ImageNet for language modeling, and estimate a similar ~0.5 OOMs/year of algorithmic efficiency trend in LLMs from 2012 to 2023. (This has wider error bars though, and doesn't capture some more recent gains, since the leading labs have stopped publishing their algorithmic efficiencies.)

Efficiency doubles roughly every 8 months

Effective compute (relative to 2014)

Figure 14: Estimates by Epoch AI of algorithmic efficiencies in language modeling. Their estimates suggest we've made ~4 OOMs of efficiency gains in 8 years.
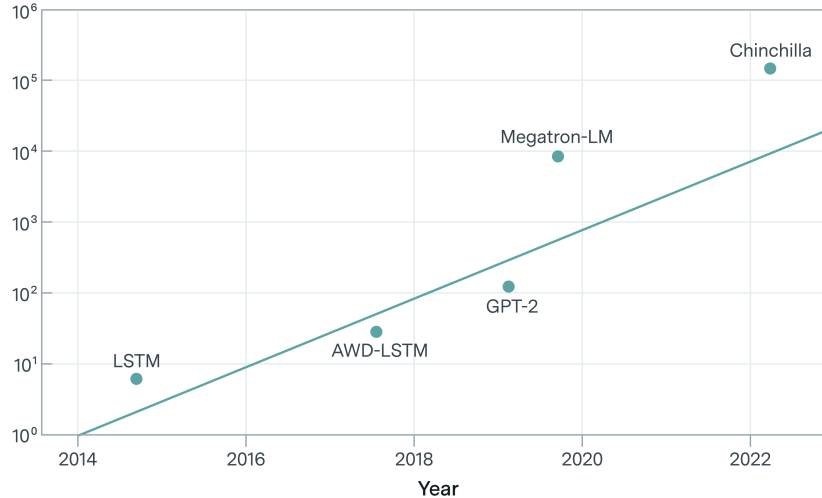
More directly looking at the last 4 years, GPT-2 to GPT-3 was basically a simple scaleup (according to the paper), but there have been many publicly-known and publicly-inferable gains since GPT-3:

- We can infer gains from API costs:[12]

  - GPT-4, on release, cost ~the same as GPT-3 when it was released, despite the absolutely enormous performance increase.[13] (If we do a naive and oversimplified back-of-the-envelope estimate based on scaling laws, this suggests that perhaps roughly half the effective compute increase from GPT-3 to GPT-4 came from algorithmic improvements.[14])

  - Since the GPT-4 release a year ago, OpenAI prices for GPT-4-level models have fallen another 6x/4x (input/output) with the release of GPT-4o.

[12] Though these are inference efficiencies (rather than necessarily training efficiencies), and to some extent will reflect inference-specific optimizations, a) they suggest enormous amounts of algorithmic progress is possible and happening in general, and b) it's often the case that an algorithmic improvements is both a training efficiency gain and an inference efficiency, for example by reducing the number of parameters necessary.

As of today (6-Dec-2024), GPT-4o costs only $2.5/1M input tokes, and $10.00/1M output tokens. So, 12x/6x cheaper than GPT-4 price mentioned below. Cost further drops by half for Batch API.

[13] GPT-3: $60/1M tokens, GPT-4: $30/1M input tokens and $60/1M output tokens.

[14] Chinchilla scaling laws say that one should scale parameter count and data equally. That is, parameter count grows "half the OOMs" of the OOMs that effective training compute grows. At the same time, parameter count is intuitively roughly proportional to inference costs. All else equal, constant inference costs thus implies that half of the OOMs of effective compute growth were "canceled out" by algorithmic win.

That said, to be clear, this is a very naive calculation (just meant for a rough illustration) that is wrong in various ways. There may be inference-specific optimizations (that don't translate into training efficiency); there may be training efficiencies that don't reduce parameter count (and thus don't translate into inference efficiency); and so on.

– Gemini 1.5 Flash, recently released, offers between "GPT-3.75-level" and GPT-4-level performance,[15] while costing 85x/57x (input/output) less than the original GPT-4 (extraordinary gains!).

- Chinchilla scaling laws give a 3x+ (0.5 OOMs+) efficiency gain.[16]

- Gemini 1.5 Pro claimed major compute efficiency gains (out-performing Gemini 1.0 Ultra, while using "significantly less" compute), with Mixture of Experts (MoE) as a highlighted architecture change. Other papers also claim a substantial multiple on compute from MoE.

- There have been many tweaks and gains on architecture, data, training stack, etc. all the time.[17]

Put together, public information suggests that the GPT-2 to GPT-4 jump included 1-2 OOMs of algorithmic efficiency gains.[18]

Figure 15: Decomposing progress: compute and algorithmic efficiencies. (Rough illustration.)
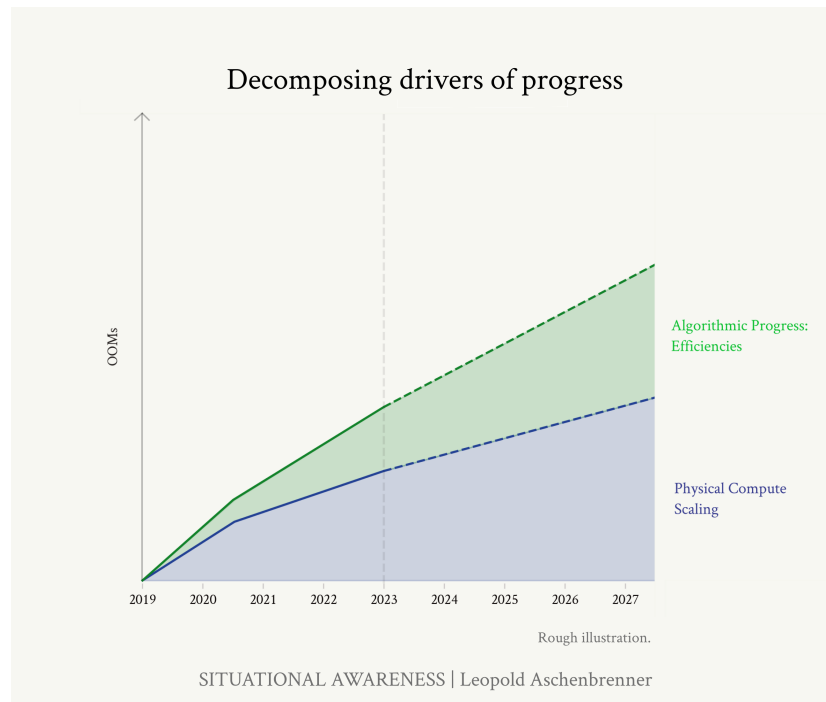
[15] Gemini 1.5 Flash ranks similarly to GPT-4 (higher than original GPT-4, lower than updated versions of GPT-4) on LMSys, a chatbot leaderboard, and has similar performance on MATH and GPQA (evals that measure reasoning) as the original GPT-4, while landing roughly in the middle between GPT-3.5 and GPT-4 on MMLU (an eval that more heavily weights towards measuring knowledge).

[16] At ~GPT-3 scale, more than 3x at larger scales.

[17] For example, this paper contains a comparison of a GPT-3-style vanilla Transformer to various simple changes to architecture and training recipe published over the years (RMSnorms instead of layernorm, different positional embeddings, SwiGlu activation, AdamW optimizer instead of Adam, etc.), what they call "Transformer++", implying a 6x gain at least at small scale.

[18] If we take the trend of 0.5 OOMs/year, and 4 years between GPT-2 and GPT-4 release, that would be 2 OOMs. However, GPT-2 to GPT-3 was a simple scaleup (after big gains from e.g. Transformers), and OpenAI claims GPT-4 pretraining finished in 2022, which could mean we're looking at closer to 2 years worth of algorithmic progress that we should be counting here. 1 OOM of algorithmic efficiency seems like a conservative lower bound.

Over the 4 years following GPT-4, we should expect the trend
to continue:[19] on average ~0.5 OOMs/year of compute effi-
ciency, i.e. ~2 OOMs of gains compared to GPT-4 by 2027.
While compute efficiencies will become harder to find as we
pick the low-hanging fruit, AI lab investments in money and
talent to find new algorithmic improvements are growing
rapidly. [20] (The publicly-inferable inference cost efficiencies,
at least, don't seem to have slowed down at all.) On the high
end, we could even see more fundamental, Transformer-like[21]
breakthroughs with even bigger gains.

Put together, this suggests we should expect something like 1-3
OOMs of algorithmic efficiency gains (compared to GPT-4) by
the end of 2027, maybe with a best guess of ~2 OOMs.

The data wall

There is a potentially important source of variance for all of
this: we're running out of internet data. That could mean
that, very soon, the naive approach to pretraining larger
language models on more scraped data could start hitting
serious bottlenecks.

Frontier models are already trained on much of the inter-
net. Llama 3, for example, was trained on over 15T tokens.
Common Crawl, a dump of much of the internet used for
LLM training, is >100T tokens raw, though much of that is
spam and duplication (e.g., a relatively simple deduplica-
tion leads to 30T tokens, implying Llama 3 would already
be using basically all the data). Moreover, for more specific
domains like code, there are many fewer tokens still, e.g.
public github repos are estimated to be in low trillions of
tokens.

You can go somewhat further by repeating data, but aca-
demic work on this suggests that repetition only gets you
so far, finding that after 16 epochs (a 16-fold repetition),
returns diminish extremely fast to nil. At some point, even
with more (effective) compute, making your models bet-
ter can become much tougher because of the data con-
straint. This isn't to be understated: we've been riding the

[19] At the very least, given over a decade
of consistent algorithmic improvements,
the burden of proof would be on those
who would suggest it will all suddenly
come to a halt!

[20] The economic returns to a 3x compute
efficiency will be measured in the $10s
of billions or more, given cluster costs.

[21] Very roughly something like a ~10x
gain.

scaling curves, riding the wave of the language-modeling-pretraining-paradigm, and without something new here, this paradigm will (at least naively) run out. Despite the massive investments, we'd plateau.

All of the labs are rumored to be making massive research bets on new algorithmic improvements or approaches to get around this. Researchers are purportedly trying many strategies, from synthetic data to self-play and RL approaches. Industry insiders seem to be very bullish: Dario Amodei (CEO of Anthropic) recently said on a podcast: "if you look at it very naively we're not that far from running out of data [. . .] My guess is that this will not be a blocker [. . .] There's just many different ways to do it." Of course, any research results on this are proprietary and not being published these days.

In addition to insider bullishness, I think there's a strong intuitive case for why it should be possible to find ways to train models with much better sample efficiency (algorithmic improvements that let them learn more from limited data). Consider how you or I would learn from a really dense math textbook:

- What a modern LLM does during training is, essentially, very very quickly skim the textbook, the words just *flying by*, not spending much brain power on it.

- Rather, when you or I read that math textbook, we read a couple pages slowly; then have an internal monologue about the material in our heads and talk about it with a few study-buddies; read another page or two; then try some practice problems, fail, try them again in a different way, get some feedback on those problems, try again until we get a problem right; and so on, until eventually the material "clicks."

- You or I also wouldn't learn much at all from a pass through a dense math textbook if all we could do was breeze through it like LLMs.²²

[22] And just rereading the same textbook over and over again might result in memorization, not understanding. I take it that's how many wordcels pass math classes!

- But perhaps, then, there are ways to incorporate aspects of how humans would digest a dense math textbook to let the models learn much more from limited data. In a simplified sense, this sort of thing—having an internal monologue about material, having a discussion with a study-buddy, trying and failing at problems until it clicks—is what many synthetic data/self-play/RL approaches are trying to do.[23]

The old state of the art of training models was simple and naive, but it worked, so nobody really tried hard to crack these approaches to sample efficiency. Now that it may become more of a constraint, we should expect all the labs to invest billions of dollars and their smartest minds into cracking it. A common pattern in deep learning is that it takes *a lot* of effort (and many failed projects) to get the details right, but eventually some version of the obvious and simple thing just works. Given how deep learning has managed to crash through every supposed wall over the last decade, my base case is that it will be similar here.

Moreover, it actually seems possible that cracking one of these algorithmic bets like synthetic data could dramatically *improve* models. Here's an intuition pump. Current frontier models like Llama 3 are trained on the internet—and the internet is mostly crap, like e-commerce or SEO or whatever. Many LLMs spend the vast majority of their training compute on this crap, rather than on really high-quality data (e.g. reasoning chains of people working through difficult science problems). Imagine if you could spend GPT-4-level compute on entirely extremely high-quality data—it could be a much, much more capable model.

A look back at AlphaGo—the first AI system that beat the world champions at Go, decades before it was thought possible—is useful here as well.[24]

- In step 1, AlphaGo was trained by imitation learning on expert human Go games. This gave it a foundation.

[23] One other way of thinking about it I find interesting: there is a "missing-middle" between pretraining and in-context learning.

In-context learning is *incredible* (and competitive with human sample efficiency). For example, the Gemini 1.5 Pro paper discusses giving the model instructional materials (a textbook, a dictionary) on Kalamang, a language spoken by fewer than 200 people and basically not present on the internet, in context—and the model learns to translate from English to Kalamang at human-level! In context, the model is able to learn from the textbook as well as a human could (and much better than it would learn from just chucking that one textbook into pretraining).

When a human learns from a textbook, they're able to distill their short-term memory/learnings into long-term memory/long-term skills with practice; however, we don't have an equivalent way to distill in-context learning "back to the weights." Synthetic data/self-play/RL/etc are trying to fix that: let the model learn by itself, then think about it and practice what it learned, distilling that learning back into the weights.

[24] See also Andrej Karpathy's talk discussing this here.

- In step 2, AlphaGo played millions of games against itself. This let it become superhuman at Go: remember the famous move 37 in the game against Lee Sedol, an extremely unusual but brilliant move a human would never have played.

Developing the equivalent of step 2 for LLMs is a key research problem for overcoming the data wall (and, moreover, will ultimately be the key to surpassing human-level intelligence).

All of this is to say that data constraints seem to inject large error bars either way into forecasting the coming years of AI progress. There's a very real chance things stall out (LLMs might still be as big of a deal as the internet, but we wouldn't get to truly crazy AGI). But I think it's reasonable to guess that the labs will crack it, and that doing so will not just keep the scaling curves going, but possibly enable huge gains in model capability.

As an aside, this also means that we should expect more variance between the different labs in coming years compared to today. Up until recently, the state of the art techniques were published, so everyone was basically doing the same thing. (And new upstarts or open source projects could easily compete with the frontier, since the recipe was published.) Now, key algorithmic ideas are becoming increasingly proprietary. I'd expect labs' approaches to diverge much more, and some to make faster progress than others—even a lab that seems on the frontier now could get stuck on the data wall while others make a breakthrough that lets them race ahead. And open source will have a much harder time competing. It will certainly make things interesting. (And if and when a lab figures it out, their breakthrough will be the key to AGI, key to superintelligence—one of the United States' most prized secrets.)

Unhobbling

Finally, the hardest to quantify—but no less important—category of improvements: what I'll call "unhobbling."

Imagine if when asked to solve a hard math problem, you had to instantly answer with the very first thing that came to mind. It seems obvious that you would have a hard time, except for the simplest problems. But until recently, that's how we had LLMs solve math problems. Instead, most of us work through the problem step-by-step on a scratchpad, and are able to solve much more difficult problems that way. "Chain-of-thought" prompting unlocked that for LLMs. Despite excellent raw capabilities, they were much worse at math than they could be because they were hobbled in an obvious way, and it took a small algorithmic tweak to unlock much greater capabilities.

We've made huge strides in "unhobbling" models over the past few years. These are algorithmic improvements beyond just training better base models—and often only use a fraction of pretraining compute—that unleash model capabilities:

- *Reinforcement learning from human feedback (RLHF)*. Base models have incredible *latent* capabilities,[25] but they're raw and incredibly hard to work with. While the popular conception of RLHF is that it merely censors swear words, RLHF has been key to making models actually useful and commercially valuable (rather than making models predict random internet text, get them to actually apply their capabilities to try to answer your question!). This was the magic of ChatGPT—well-done RLHF made models usable and useful to real people for the first time. The original InstructGPT paper has a great quantification of this: an RLHF'd small model was equivalent to a non-RLHF'd >100x larger model in terms of human rater preference.

- *Chain of Thought (CoT)*. As discussed. CoT started being widely used just 2 years ago and can provide the equivalent of a >10x effective compute increase on math/reasoning problems.

- *Scaffolding*. Think of CoT++: rather than just asking a model

[25] That's the magic of unsupervised learning, in some sense: to better predict the next token, to make perplexity go down, models learn incredibly rich internal representations, everything from (famously) sentiment to complex world models. But, out of the box, they're hobbled: they're using their incredible internal representations merely to predict the next token in random internet text, and rather than applying them in the best way to actually try to solve your problem.

to solve a problem, have one model make a plan of attack, have another propose a bunch of possible solutions, have another critique it, and so on. For example, on HumanEval (coding problems), simple scaffolding enables GPT-3.5 to outperform un-scaffolded GPT-4. On SWE-Bench (a benchmark of solving real-world software engineering tasks), GPT-4 can only solve ~2% correctly, while with Devin's agent scaffolding it jumps to 14-23%. (Unlocking agency is only in its infancy though, as I'll discuss more later.)

- *Tools*: Imagine if humans weren't allowed to use calculators or computers. We're only at the beginning here, but ChatGPT can now use a web browser, run some code, and so on.

- *Context length*. Models have gone from 2k token context (GPT-3) to 32k context (GPT-4 release) to 1M+ context (Gemini 1.5 Pro). This is a *huge* deal. A much smaller base model with, say, 100k tokens of relevant context can outperform a model that is much larger but only has, say, 4k relevant tokens of context—more context is effectively a large compute efficiency gain.[26] More generally, context is key to unlocking many applications of these models: for example, many coding applications require understanding large parts of a codebase in order to usefully contribute new code; or, if you're using a model to help you write a document at work, it really needs the context from lots of related internal docs and conversations. Gemini 1.5 Pro, with its 1M+ token context, was even able to learn a new language (a low-resource language not on the internet) from scratch, just by putting a dictionary and grammar reference materials in context!

- *Posttraining improvements*. The current GPT-4 has substantially improved compared to the original GPT-4 when released, according to John Schulman due to posttraining improvements that unlocked latent model capability: on reasoning evals it's made substantial gains (e.g., ~50% -> 72% on MATH, ~40% to ~50% on GPQA) and on the LMSys leaderboard, it's made nearly 100-point elo jump (comparable to the difference in elo between Claude 3 Haiku and the much larger Claude 3 Opus, models that have a ~50x price difference).

[26] See Figure 7 from the updated Gemini 1.5 whitepaper, comparing perplexity vs. context for Gemini 1.5 Pro and Gemini 1.5 Flash (a much cheaper and presumably smaller model).

What is a post-training improvement?

A survey by Epoch AI of some of these techniques, like scaffolding, tool use, and so on, finds that techniques like this can typically result in effective compute gains of 5-30x on many benchmarks. METR (an organization that evaluates models) similarly found very large performance improvements on their set of agentic tasks, via unhobbling from the same GPT-4 base model: from 5% with just the base model, to 20% with the GPT-4 as posttrained on release, to nearly 40% today from better posttraining, tools, and agent scaffolding (Figure 16).
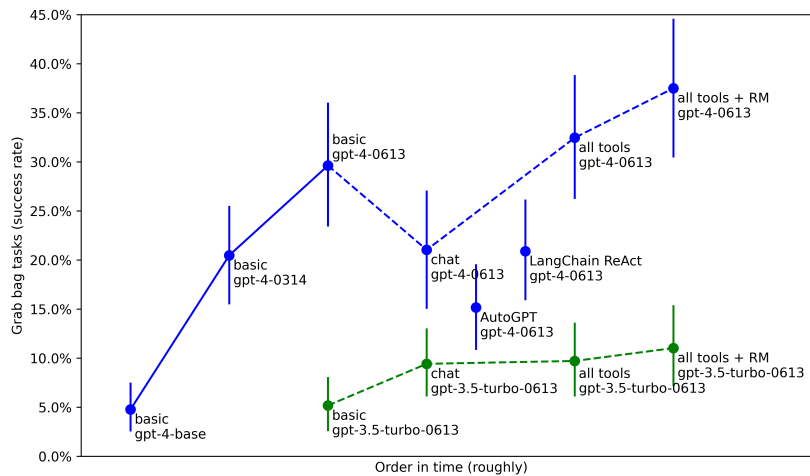
Figure 16: Performance on METR's agentic tasks, over time via better unhobbling. Source: Model Evaluation and Threat Research

While it's hard to put these on a unified effective compute scale with compute and algorithmic efficiencies, it's clear these are *huge* gains, at least on a roughly similar magnitude as the compute scaleup and algorithmic efficiencies. (It also highlights the central role of algorithmic progress: the ~0.5 OOMs/year of compute efficiencies, already significant, are only part of the story, and put together with unhobbling algorithmic progress overall is maybe even a majority of the gains on the current trend.)

"Unhobbling" is a huge part of what actually enabled these models to become useful—and I'd argue that much of what is holding back many commercial applications today is the need for further "unhobbling" of this sort. Indeed, *models today are still incredibly hobbled*! For example:

- They don't have long-term memory.
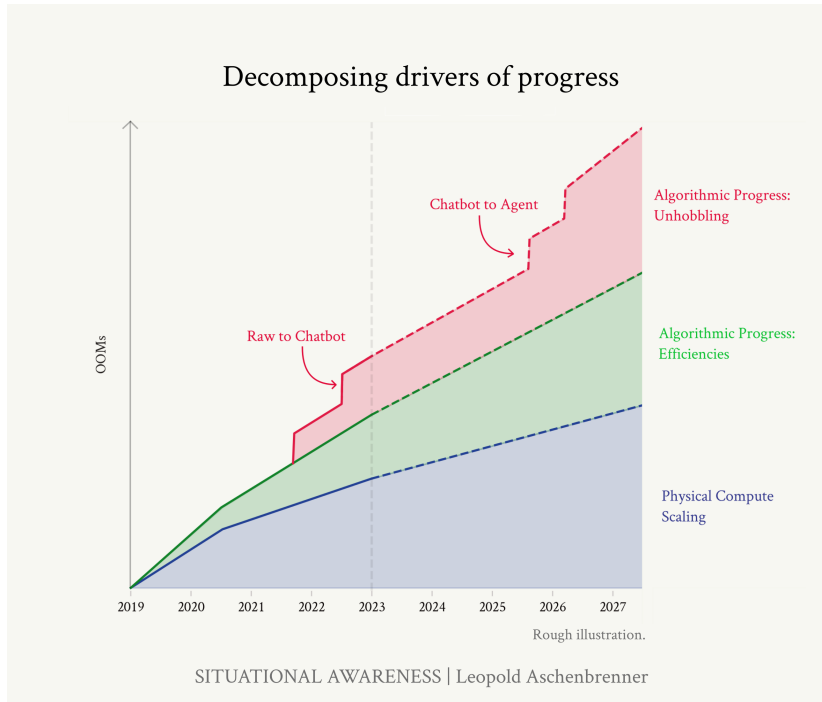
- They can't use a computer (they still only have very limited tools).

- They still mostly don't think before they speak. When you ask ChatGPT to write an essay, that's like expecting a human to write an essay via their initial stream-of-consciousness.[27]

- They can (mostly) only engage in short back-and-forth dialogues, rather than going away for a day or a week, thinking about a problem, researching different approaches, consulting other humans, and then writing you a longer report or pull request.

- They're mostly not personalized to you or your application (just a generic chatbot with a short prompt, rather than having all the relevant background on your company and your work).

The possibilities here are enormous, and we're rapidly picking low-hanging fruit here. This is critical: *it's completely wrong to just imagine "GPT-6 ChatGPT."* With continued unhobbling

[27] People are working on this though.

progress, the improvements will be step-changes compared to GPT-6 + RLHF. By 2027, rather than a chatbot, you're going to have something that looks more like an agent, like a coworker.

From chatbot to agent-coworker

What could ambitious unhobbling over the coming years look like? The way I think about it, there are three key ingredients:

1. SOLVING THE "ONBOARDING PROBLEM"

GPT-4 has the raw smarts to do a decent chunk of many people's jobs, but it's sort of like a smart new hire that just showed up 5 minutes ago: it doesn't have any relevant context, hasn't read the company docs or Slack history or had conversations with members of the team, or spent any time understanding the company-internal codebase. A smart new hire isn't that useful 5 minutes after arriving—but they are quite useful a month in! It seems like it should be possible, for example via very-long-context, to "onboard" models like we would a new human coworker. This alone would be a huge unlock.

2. THE TEST-TIME COMPUTE OVERHANG (REASONING/ERROR CORRECTION/SYSTEM II FOR LONGER-HORIZON PROBLEMS)

Right now, models can basically only do short tasks: you ask them a question, and they give you an answer. But that's extremely limiting. Most useful cognitive work humans do is longer horizon—it doesn't just take 5 minutes, but hours, days, weeks, or months.

A scientist that could only think about a difficult problem for 5 minutes couldn't make any scientific breakthroughs. A software engineer that could only write skeleton code for a single function when asked wouldn't be very useful— software engineers are given a larger task, and they then go make a plan, understand relevant parts of the codebase or

technical tools, write different modules and test them incrementally, debug errors, search over the space of possible solutions, and eventually submit a large pull request that's the culmination of weeks of work. And so on.

In essence, there is a large *test-time compute overhang.* Think of each GPT-4 token as a word of internal monologue when you think about a problem. Each GPT-4 token is quite smart, but it can currently only really effectively use on the order of ~hundreds of tokens for chains of thought coherently (effectively as though you could only spend a few minutes of internal monologue/thinking on a problem or project).

What if it could use millions of tokens to think about and work on really hard problems or bigger projects?

Number of tokens	Equivalent to me working on something for...	
100s	A few minutes	ChatGPT (we are here)
1,000s	Half an hour	+1 OOMs test-time compute
10,000s	Half a workday	+2 OOMs
100,000s	A workweek	+3 OOMs
Millions	Multiple months	+4 OOMs

Table 2: Assuming a human thinking at ~100 tokens/minute and working 40 hours/week, translating "how long a model thinks" in tokens to human-time on a given problem/project.

Even if the "per-token" intelligence were the same, it'd be the difference between a smart person spending a *few minutes* vs. a *few months* on a problem. I don't know about you, but there's much, much, much more I am capable of in a few months vs. a few minutes. If we could unlock "being able to think and work on something for months-equivalent, rather than a few-minutes-equivalent" for models, it would unlock an *insane* jump in capability. There's a huge overhang here, many OOMs worth.

Right now, models can't do this yet. Even with recent advances in long-context, this longer context mostly only works for the consumption of tokens, not the production of tokens—after a while, the model goes off the rails or gets stuck. It's not yet able to go away for a while to work on a

problem or project on its own.[28]

But unlocking test-time compute might merely be a matter of relatively small "unhobbling" algorithmic wins. Perhaps a small amount of RL helps a model learn to error correct ("hm, that doesn't look right, let me double check that"), make plans, search over possible solutions, and so on. In a sense, the model already has most of the raw capabilities, it just needs to learn a few extra skills on top to put it all together.

In essence, we just need to teach the model a sort of System II outer loop[29] that lets it reason through difficult, long-horizon projects.

If we succeed at teaching this outer loop, instead of a short chatbot answer of a couple paragraphs, imagine a stream of millions of words (coming in more quickly than you can read them) as the model thinks through problems, uses tools, tries different approaches, does research, revises its work, coordinates with others, and completes big projects on its own.

Trading off test-time and train-time compute in other ML domains.   In other domains, like AI systems for board games, it's been demonstrated that you can use more test-time compute (also called inference-time compute) to substitute for training compute.
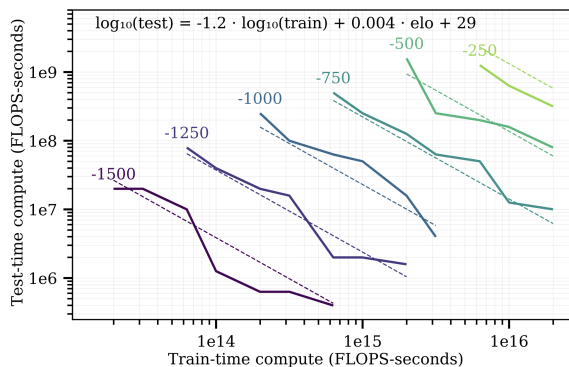
Fig. 9. The trade-off between train-time compute and test-time compute. Each dotted line gives the minimum train-test compute required for a certain Elo on a 9 × 9 board

[28] Which makes sense—why would it have learned the skills for longer-horizon reasoning and error correction? There's very little data on the internet in the form of "my complete internal monologue, reasoning, all the relevant steps over the course of a month as I work on a project." Unlocking this capability will require a new kind of training, for it to learn these extra skills.
   Or as Gwern put it (private correspondence): " 'Brain the size of a galaxy, and what do they ask me to do? Predict the misspelled answers on benchmarks!' Marvin the depressed neural network moaned."

[29] System I vs. System II is a useful way of thinking about current capabilities of LLMs—including their limitations and dumb mistakes—and what might be possible with RL and unhobbling. Think of this way: when you are driving, most of the time you are on autopilot (system I, what models mostly do right now). But when you encounter a complex construction zone or novel intersection, you might ask your passenger-seat-companion to pause your conversation for a moment while you figure out—actually think about—what's going on and what to do. If you were forced to go about life with only system I (closer to models today), you'd have a lot of trouble. Creating the ability for system II reasoning loops is a central unlock.

Figure 18: Jones (2021): A smaller model can do as well as a much larger model at the game of Hex if you give it more test-time compute ("more time to think"). In this domain, they find that one can spend ~1.2 OOMs more compute at test-time to get performance equivalent to a model with ~1 OOMs more training compute.

If a similar relationship held in our case, if we could unlock +4 OOMs of test-time compute, that might be equivalent to +3 OOMs of pretraining compute, i.e. very roughly something like the jump between GPT-3 and GPT-4. (Solving this "unhobbling" would be equivalent to a huge OOM scaleup.)

3. USING A COMPUTER

This is perhaps the most straightforward of the three. Chat-GPT right now is basically like a human that sits in an isolated box that you can text. While early unhobbling improvements teach models to use individual isolated tools, I expect that with multimodal models we will soon be able to do this in one fell swoop: we will simply enable models to use a computer like a human would.

That means joining your Zoom calls, researching things online, messaging and emailing people, reading shared docs, using your apps and dev tooling, and so on. (Of course, for models to make the most use of this in longer-horizon loops, this will go hand-in-hand with unlocking test-time compute.)

BY THE END OF THIS, I expect us to get something that looks a lot like a *drop-in remote worker*. An agent that joins your company, is onboarded like a new human hire, messages you and colleagues on Slack and uses your softwares, makes pull requests, and that, given big projects, can do the model-equivalent of a human going away for weeks to independently complete the project. You'll probably need somewhat better base models than GPT-4 to unlock this, but possibly not even *that* much better—a lot of juice is in fixing the clear and basic ways models are still hobbled.

(A very early peek at what this might look like is Devin, an early prototype of unlocking the "agency-overhang"/"test-time compute overhang" on models on the path to creating a fully automated software engineer. I don't know how well Devin works in practice, and this demo is still very

limited compared to what proper chatbot → agent un-hobbling would yield, but it's a useful teaser of the sort of thing coming soon.)

By the way, the centrality of unhobbling might lead to a some-what interesting "sonic boom" effect in terms of commercial applications. Intermediate models between now and the drop-in remote worker will require tons of schlep to change work-flows and build infrastructure to integrate and derive economic value from. The drop-in remote worker will be dramatically easier to integrate—just, well, drop them in to automate all the jobs that could be done remotely. It seems plausible that the schlep will take longer than the unhobbling, that is, by the time the drop-in remote worker is able to automate a large number of jobs, intermediate models won't yet have been fully harnessed and integrated—so the jump in economic value gen-erated could be somewhat discontinuous.

What does author mean by "schelp" vs. "unhobbling"?

The next four years

PUTTING THE NUMBERS TOGETHER, we should (roughly) ex-pect another GPT-2-to-GPT-4-sized jump in the 4 years follow-ing GPT-4, by the end of 2027.

- GPT-2 to GPT-4 was roughly a 4.5–6 OOM base effective compute scaleup (physical compute and algorithmic efficien-cies), plus major "unhobbling" gains (from base model to chatbot).

- In the subsequent 4 years, we should expect 3–6 OOMs of base effective compute scaleup (physical compute al-gorithmic efficiencies)—with perhaps a best guess of ~5 OOMs—plus step-changes in utility and applications un-locked by "unhobbling" (from chatbot to agent/drop-in remote worker).

To put this in perspective, suppose GPT-4 training took 3 months. *In 2027, a leading AI lab will be able to train a GPT-4-level model in a minute.*[30] The OOM effective compute scaleup

[30] On the best guess assumptions on physical compute and algorithmic efficiency scaleups described above, and simplifying parallelism considerations (in reality, it might look more like "1440 (60*24) GPT-4-level models in a day" or similar).

Minutes taken to train a GPT model now = 3*90*24*60 = 3,88,880
With ~5OOMs improvement over next 4 yrs, it would just take 3,88,880/1e5 = 3.8 mins

GPT-2 (2019) to GPT-4 (2024)

Compute	Algorithmic Efficiency	Unhobbling
3.5-4 OOMs	1-2 OOMs	2? OOMs RLHF / CoT / Scaffolding / Basic tools / Etc.

4.5-6 OOMs of base scaleup · Base to Chatbot

Based on public estimates.

SITUATIONAL AWARENESS | Leopold Aschenbrenner

2023-2027 (Projection)

Compute	Algorithmic Efficiency	Unhobbling
2-3 OOMs	1-3 OOMs	? OOMs Onboarding problem / Test-time compute/ System II / Using a computer

3-6 OOMs (best guess: ~5 OOMs) of base scaleup · Chatbot to Agent

Based on public estimates.

SITUATIONAL AWARENESS | Leopold Aschenbrenner

Figure 19: Summary of the estimates on drivers of progress in the four years preceding GPT-4, and what we should expect in the four years following GPT-4.

will be dramatic.

Where will that take us?

Counting the OOMs

| Model | GPT-2 (2019) | GPT-3 (2020) | GPT-4 (2023) | 2027 |

OOMs: 0 1 2 3 4 5 6 7 8 9 10 11

Intelligence: Preschooler — Elementary Schooler — Smart High Schooler — Automated AI Researcher/Engineer?

Unhobbling: Base — Chatbot — Agent

SITUATIONAL AWARENESS | Leopold Aschenbrenner
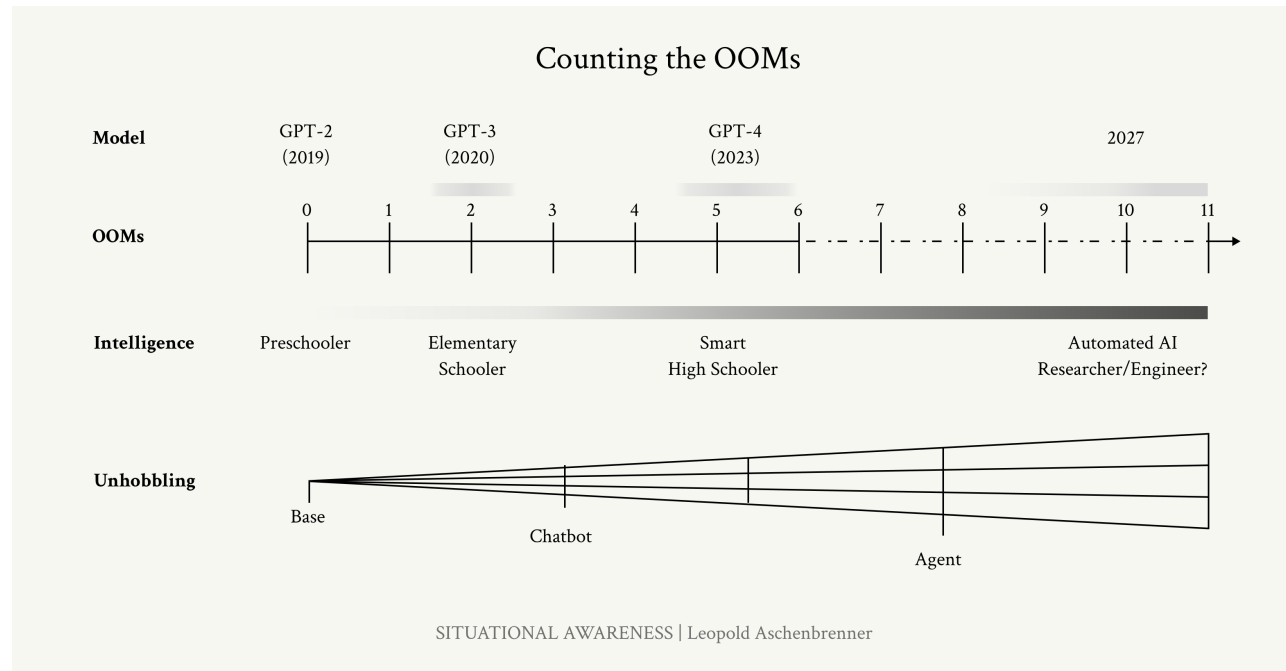
Figure 20: ==Summary of counting the OOMs==.

GPT-2 to GPT-4 took us from ~preschooler to ~smart high-schooler; from barely being able to output a few cohesive sentences to acing high-school exams and being a useful coding assistant. That was an insane jump. If *this* is the intelligence gap we'll cover once more, where will that take us?[31] We should not be surprised if that takes us very, very far. ==Likely, it will take us to models that can outperform PhDs and the best experts in a field.==

(One neat way to think about this is that the ==current trend of AI progress is proceeding at roughly 3x the pace of child development==. Your 3x-speed-child just graduated high school; it'll be taking your job before you know it!)

Again, critically, don't just imagine an incredibly smart Chat-GPT: unhobbling gains should mean that this looks more like a drop-in remote worker, an incredibly smart agent that can reason and plan and error-correct and knows everything about you and your company and can work on a problem indepen-

[31] Of course, any benchmark we have today will be saturated. But that's not saying much; it's mostly a reflection on the difficulty of making hard-enough benchmarks.

dently for weeks.

We are on course for AGI by 2027. These AI systems will basically be able to automate basically all cognitive jobs (think: all jobs that could be done remotely).

To be clear—the error bars are large. Progress could stall as we run out of data, if the algorithmic breakthroughs necessary to crash through the data wall prove harder than expected. Maybe unhobbling doesn't go as far, and we are stuck with merely expert chatbots, rather than expert coworkers. Perhaps the decade-long trendlines break, or scaling deep learning hits a wall for real this time. (Or an algorithmic breakthrough, even simple unhobbling that unleashes the test-time compute overhang, could be a paradigm-shift, accelerating things further and leading to AGI even earlier.)

In any case, we are racing through the OOMs, and it requires no esoteric beliefs, merely trend extrapolation of straight lines, to take the possibility of AGI—true AGI—by 2027 *extremely* seriously.

It seems like many are in the game of downward-defining AGI these days, as just as really good chatbot or whatever. What I mean is an AI system that could fully automate my or my friends' job, that could fully do the work of an AI researcher or engineer. Perhaps some areas, like robotics, might take longer to figure out by default. And the societal rollout, e.g. in medical or legal professions, could easily be slowed by societal choices or regulation. But once models can automate AI research itself, that's enough—enough to kick off intense feedback loops—and we could very quickly make further progress, the automated AI engineers themselves solving all the remaining bottlenecks to fully automating everything. In particular, millions of automated researchers could very plausibly compress a decade of further algorithmic progress into a year or less. AGI will merely be a small taste of the superintelligence soon to follow. (More on that in the next chapter.)

In any case, do not expect the vertiginous pace of progress to abate. The trendlines look innocent, but their implications are

intense. As with every generation before them, every new generation of models will dumbfound most onlookers; they'll be incredulous when, very soon, models solve incredibly difficult science problems that would take PhDs days, when they're whizzing around your computer doing your job, when they're writing codebases with millions of lines of code from scratch, <mark>when every year or two the economic value generated by these models 10xs</mark>. Forget scifi, count the OOMs: it's what we should expect. AGI is no longer a distant fantasy. Scaling up simple deep learning techniques has just worked, the models just want to learn, and we're about to do another 100,000x+ by the end of 2027. It won't be long before they're smarter than us.

2014
2015
2016
2017
2018

Figure 21: GPT-4 is just the beginning—where will we be four years later? Do not make the mistake of underestimating the rapid pace of deep learning progress (as illustrated by progress in GANs).

Addendum. Racing through the OOMs: It's this decade or bust

I used to be more skeptical of short timelines to AGI. One reason is that it seemed unreasonable to privilege this decade, concentrating so much AGI-probability-mass on it (it seemed like a classic fallacy to think "oh we're so special"). I thought we should be uncertain about what it takes to get AGI, which should lead to a much more "smeared-out" probability distribution over when we might get AGI.

However, I've changed my mind: critically, our uncertainty over what it takes to get AGI should be over *OOMs* (of effective compute), rather than over years.

We're racing through the OOMs this decade. Even at its bygone heyday, Moore's law was only 1–1.5 OOMs/decade. I estimate that we will do ~5 OOMs in 4 years, and over ~10 this decade overall.
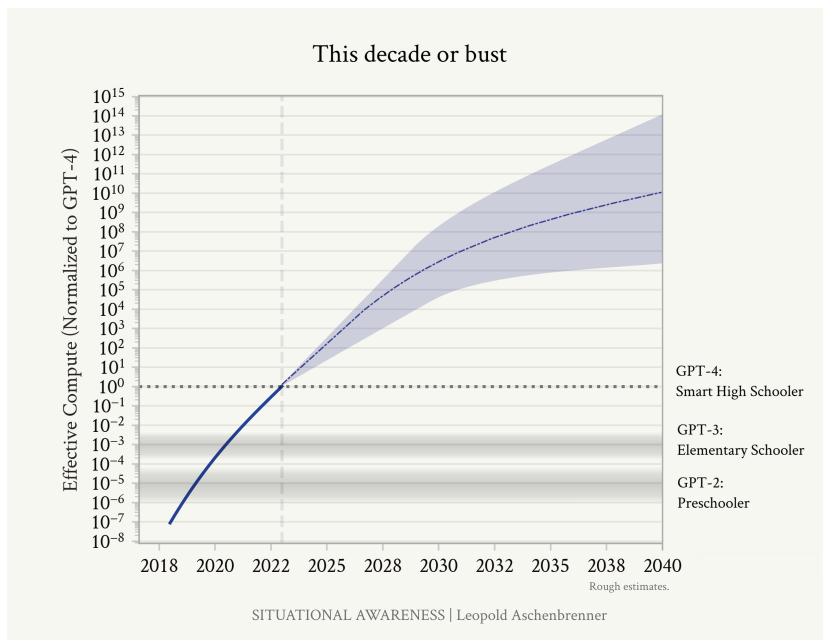
Figure 22: Rough projections on effective compute scaleup. We've been racing through the OOMs this decade; after the early 2030s, we will face a slow slog.

In essence, we're in the middle of a huge scaleup reaping one-time gains this decade, and progress through the OOMs will be multiples slower thereafter. If this scaleup doesn't get us to AGI in the next 5-10 years, it might be a

long way out.

- *Spending scaleup*: Spending a million dollars on a model used to be outrageous; by the end of the decade, we will likely have $100B or $1T clusters. Going much higher than that will be hard; that's already basically the feasible limit (both in terms of what big business can afford, and even just as a fraction of GDP). Thereafter all we have is glacial 2%/year trend real GDP growth to increase this.

- *Hardware gains*: AI hardware has been improving much more quickly than Moore's law. That's because we've been specializing chips for AI workloads. For example, we've gone from CPUs to GPUs; adapted chips for Transformers; and we've gone down to much lower precision number formats, from fp64/fp32 for traditional supercomputing to fp8 on H100s. These are large gains, but by the end of the decade we'll likely have totally-specialized AI-specific chips, without much further beyond-Moore's law gains possible.

- *Algorithmic progress*: In the coming decade, AI labs will invest tens of billions in algorithmic R&D, and all the smartest people in the world will be working on this; from tiny efficiencies to new paradigms, we'll be picking lots of the low-hanging fruit. We probably won't reach any sort of hard limit (though "unhobblings" are likely finite), but at the very least the pace of improvements should slow down, as the rapid growth (in $ and human capital investments) necessarily slows down (e.g., most of the smart STEM talent will already be working on AI). (That said, this is the most uncertain to predict, and the source of most of the uncertainty on the OOMs in the 2030s on the plot above.)

Put together, this means we are racing through many more OOMs in the next decade than we might in multiple decades thereafter. Maybe it's enough—and we get AGI soon—or we might be in for a long, slow slog. You and I can reasonably disagree on the *median* time to AGI,

depending on how hard we think achieving AGI will be—but given how we're racing through the OOMs right now, certainly your *modal* AGI year should sometime later this decade or so.

Matthew Barnett
@MatthewJBar
···

My own basic calculations suggest that, given the potential for increased investment and hardware progress, we could very soon move through a large fraction of the remaining compute gap between the current frontier models and the literal amount of computation used by evolution.

7:54 PM · Mar 26, 2024 · **3,968** Views

Figure 23: Matthew Barnett has a nice related visualization of this, considering just compute and biological bounds.

II. From AGI to Superintelligence: the Intelligence Explosion

AI progress won't stop at human-level. Hundreds of millions of AGIs could automate AI research, compressing a decade of algorithmic progress (5+ OOMs) into 1 year. We would rapidly go from human-level to vastly superhuman AI systems. The power—and the peril—of superintelligence would be dramatic.

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make.

<div align="right">

I. J. GOOD (1965)

</div>

The Bomb and The Super
In the common imagination, the Cold War's terrors principally trace back to Los Alamos, with the invention of the atomic bomb. But The Bomb, alone, is perhaps overrated. Going from The Bomb to The Super—hydrogen bombs—was arguably just as important.

In the Tokyo air raids, hundreds of bombers dropped thousands of tons of conventional bombs on the city. Later that year, Little Boy, dropped on Hiroshima, unleashed similar destructive power in a single device. But just 7 years later, Teller's hydrogen bomb multiplied yields *a thousand-fold* once again—a single bomb with more explosive power than all of the bombs dropped in the entirety of WWII combined.

The Bomb was a more efficient bombing campaign. The Super was a country-annihilating device.[32]

So it will be with AGI and Superintelligence.

AI PROGRESS WON'T STOP AT HUMAN-LEVEL. After initially learning from the best human games, AlphaGo started playing against itself—and it quickly became superhuman, playing extremely creative and complex moves that a human would never have come up with.

We discussed the path to AGI in the previous piece. Once we get AGI, we'll turn the crank one more time—or two or three more times—and AI systems will become superhuman—vastly superhuman. They will become qualitatively smarter than you or I, much smarter, perhaps similar to how you or I are qualitatively smarter than an elementary schooler.

The jump to superintelligence would be wild enough at the current rapid but continuous rate of AI progress (if we could make the jump to AGI in 4 years from GPT-4, what might another 4 or 8 years after that bring?). But it could be much faster than that, if AGI automates AI research itself.

Once we get AGI, we won't just have one AGI. I'll walk through the numbers later, but: given inference GPU fleets by then, we'll likely be able to run *many millions of them (perhaps 100 million human-equivalents, and soon after at 10x+ human speed).* Even if they can't yet walk around the office or make coffee, they will be able to do ML research on a computer. Rather than a few hundred researchers and engineers at a leading AI lab, we'd have more than 100,000x that—furiously working on algorithmic breakthroughs, day and night. Yes, recursive self-improvement, but no sci-fi required; *they would need only to accelerate the existing trendlines of algorithmic progress (currently at ~0.5 OOMs/year).*

Automated AI research could probably compress a human-decade of algorithmic progress into less than a year (and that seems conservative). That'd be 5+ OOMs, another GPT-2-to-

Scenario: Intelligence Explosion

SITUATIONAL AWARENESS | Leopold Aschenbrenner

Figure 24: Automated AI research could accelerate algorithmic progress, leading to 5+ OOMs of effective compute gains in a year. The AI systems we'd have by the end of an intelligence explosion would be vastly smarter than humans.
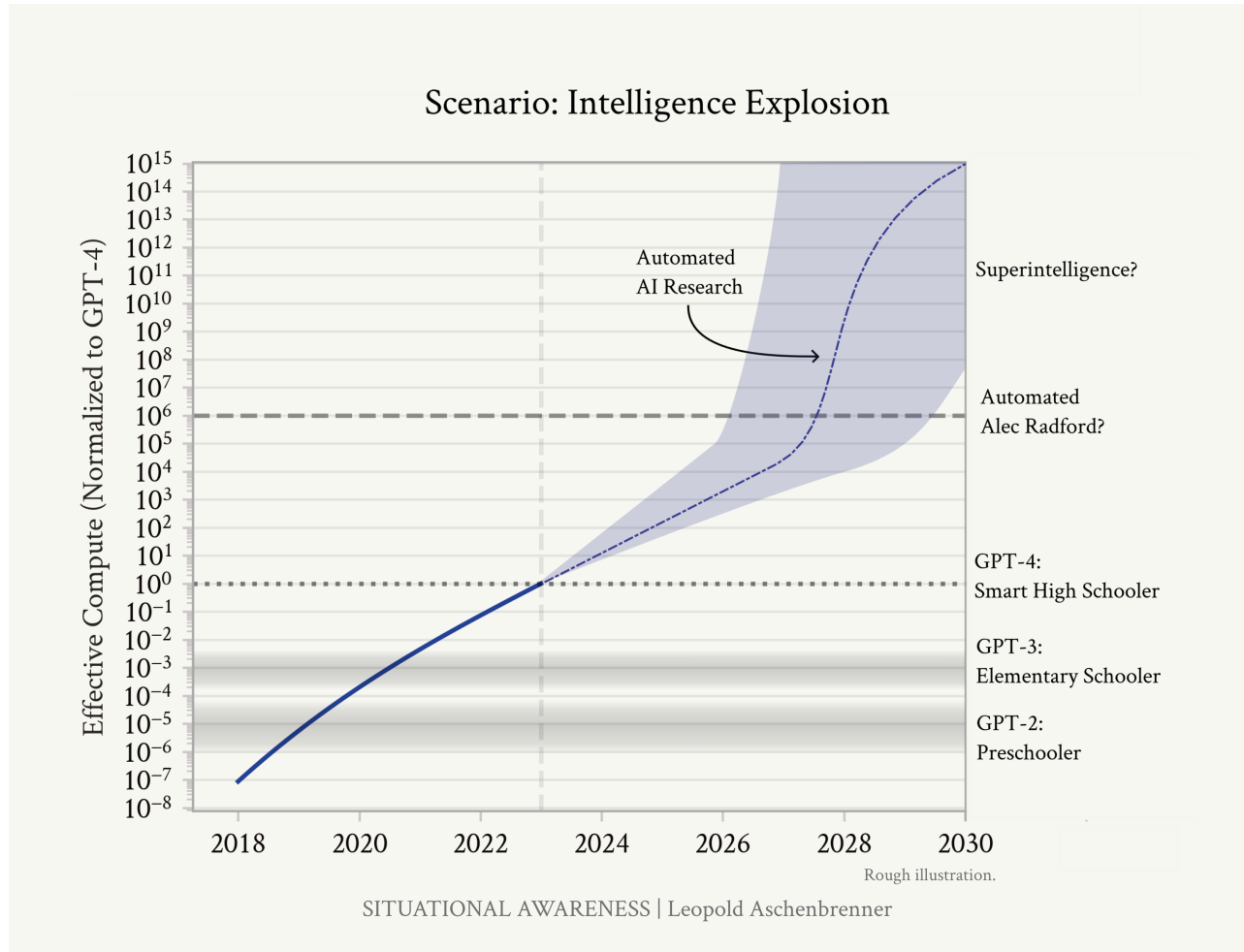
GPT-4-sized jump, on top of AGI—a qualitative jump like that from a preschooler to a smart high schooler, *on top* of AI systems already as smart as expert AI researchers/engineers.

There are several plausible bottlenecks—including limited compute for experiments, complementarities with humans, and algorithmic progress becoming harder—which I'll address, but none seem sufficient to definitively slow things down.

Before we know it, we would have superintelligence on our hands—AI systems *vastly* smarter than humans, capable of novel, creative, complicated behavior we couldn't even begin to understand—perhaps even a small civilization of billions of them. Their power would be vast, too. Applying superin-

telligence to R&D in other fields, explosive progress would broaden from just ML research; soon they'd solve robotics, make dramatic leaps across other fields of science and technology within years, and an industrial explosion would follow. Superintelligence would likely provide a decisive military advantage, and unfold untold powers of destruction. We will be faced with one of the most intense and volatile moments of human history.

Automating AI research

We don't need to automate everything—just AI research.   A common objection to transformative impacts of AGI is that it will be hard for AI to do everything. Look at robotics, for instance, doubters say; that will be a gnarly problem, even if AI is cognitively at the levels of PhDs. Or take automating biology R&D, which might require lots of physical lab-work and human experiments.

But we don't need robotics—we don't need many things—for AI to automate AI research. The jobs of AI researchers and engineers at leading labs can be done fully virtually and don't run into real-world bottlenecks in the same way (though it will still be limited by compute, which I'll address later). And the job of an AI researcher is fairly straightforward, in the grand scheme of things: read ML literature and come up with new questions or ideas, implement experiments to test those ideas, interpret the results, and repeat. This all seems squarely in the domain where simple extrapolations of current AI capabilities could easily take us to or beyond the levels of the best humans by the end of 2027.[33]

It's worth emphasizing just how straightforward and hacky some of the biggest machine learning breakthroughs of the last decade have been: "oh, just add some normalization" (LayerNorm/BatchNorm) or "do f(x)+x instead of f(x)" (residual connections)" or "fix an implementation bug" (Kaplan $\rightarrow$ Chinchilla scaling laws). AI research can be automated. And automating AI research is all it takes to kick off extraordinary feedback loops.[34]

[33] The job of an AI researcher is also a job that AI researchers at AI labs just, well, know really well—so it'll be particularly intuitive to them to optimize models to be good at that job. And there will be huge incentives to do so to help them accelerate their research and their labs' competitive edge.

[34] This suggests an important point in terms of the sequencing of risks from AI, by the way. A common AI threat model people point to is AI systems developing novel bioweapons, and that posing catastrophic risk. But if AI research is more straightforward to automate than biology R&D, we might get an intelligence explosion before we get extreme AI biothreats. This matters, for example, with regard to whether we should expect "bio warning shots" in time before things get crazy on AI.

We'd be able to run millions of copies (and soon at 10x+ human speed) of the automated AI researchers. Even by 2027, we should expect GPU fleets in the 10s of millions. Training clusters alone should be approaching ~3 OOMs larger, already putting us at 10 million+ A100-equivalents. Inference fleets should be much larger still. (More on all this in IIIa. Racing to the Trillion Dollar Cluster.)

That would let us run many millions of copies of our automated AI researchers, perhaps 100 million human-researcher-equivalents, running day and night. There's some assumptions that flow into the exact numbers, including that humans "think" at 100 tokens/minute (just a rough order of magnitude estimate, e.g. consider your internal monologue) and extrapolating historical trends and Chinchilla scaling laws on per-token inference costs for frontier models remaining in the same ballpark.[35] We'd also want to reserve some of the GPUs for running experiments and training new models. Full calculation in a footnote.[36]

Another way of thinking about it is that *given inference fleets in 2027, we should be able to generate an entire internet's worth of tokens, every single day.*[37] In any case, the exact numbers don't matter that much, beyond a simple plausibility demonstration.

[35] As noted earlier, the GPT-4 API costs less today than GPT-3 when it was released—this suggests that the trend of inference efficiency wins is fast enough to keep inference costs roughly constant even as models get much more powerful. Similarly, there have been huge inference cost wins in just the year since GPT-4 was released; for example, the current version of Gemini 1.5 Pro outperforms the original GPT-4, while being roughly 10x cheaper.

We can also ground this somewhat more by considering Chinchilla scaling laws. On Chinchilla scaling laws, model size—and thus inference costs—grow with the square root of training cost, i.e. half the OOMs of the OOM scaleup of effective compute. However, in the previous piece, I suggested that algorithmic efficiency was advancing at roughly the same pace as compute scaleup, i.e. it made up roughly half of the OOMs of effective compute scaleup. If these algorithmic wins also translate into inference efficiency, that means that the algorithmic efficiencies would compensate for the naive increase in inference cost.

In practice, training compute efficiencies often, but not always, translate into inference efficiency wins. However, there are also separately many inference efficiency wins that are not training efficiency wins. So, at least in terms of the rough ballpark, assuming the $/token of frontier models stays roughly similar doesn't seem crazy.

(Of course, they'll use more tokens, i.e. more test-time compute. But that's already part of the calculation here, by pricing human-equivalents as 100 tokens/minute.)

[36] GPT4T is about $0.03/1K tokens. We supposed we would have 10s of millions of A100 equivalents, which cost ~$1 hour per GPU if A100-equivalents. If we used the API costs to translate GPUs into tokens generated, that implies 10s of millions GPUs * $1/GPU-hour * 33K tokens/$ = ~ one trillion tokens/ hour. Suppose a human does 100 tokens/min of thinking, that means a human-equivalent is 6,000 tokens/hour. One trillion tokens/hour divided by 6,000 tokens/human-hour = ~200 million human-equivalents— i.e. as if running 200 million human researchers, day and night. (And even if we reserve half the GPUs for experiment compute, we get 100 million human-researcher-equivalents.)

[37] The previous footnote estimated ~1T tokens/hour, i.e. 24T tokens a day. In the previous piece, I noted that a public deduplicated CommonCrawl had around 30T tokens.

Moreover, our automated AI researchers may soon be able to run at much faster than human-speed:

- By taking some inference penalties, we can trade off running fewer copies in exchange for running them at faster serial speed. (For example, we could go from ~5x human speed to ~100x human speed by "only" running 1 million copies of the automated researchers.[38])

- More importantly, the first algorithmic innovation the automated AI researchers work on is getting a 10x or 100x speedup. Gemini 1.5 Flash is ~10x faster than the originally-released GPT-4,[39] merely a year later, while providing similar performance to the originally-released GPT-4 on reasoning benchmarks. If that's the algorithmic speedup a few hundred human researchers can find in a year, the automated AI researchers will be able to find similar wins very quickly.

That is: expect 100 million automated researchers each working at 100x human speed not long after we begin to be able to automate AI research. They'll each be able to do a year's worth of work in a few days. The increase in research effort—compared to a few hundred puny human researchers at a leading AI lab today, working at a puny 1x human speed—will be extraordinary.

This could easily dramatically accelerate existing trends of algorithmic progress, compressing a decade of advances into a year.   We need not postulate anything totally novel for automated AI research to intensely speed up AI progress. Walking through the numbers in the previous piece, we saw that algorithmic progress has been a central driver of deep learning progress in the last decade; we noted a trendline of ~0.5 OOMs/year on algorithmic efficiencies alone, with additional large algorithmic gains from unhobbling on top. (I think the import of algorithmic progress has been underrated by many, and properly appreciating it is important for appreciating the possibility of an intelligence explosion.)

Could our millions of automated AI researchers (soon working at 10x or 100x human speed) compress the algorithmic progress human researchers would have found in a decade into a year

[38] Jacob Steinhardt estimates that k^3 parallel copies of a model can be replaced with a single model that is k^2 faster, given some math on inference tradeoffs with a tiling scheme (that theoretically works even for k of 100 or more). Suppose initial speeds were already ~5x human speed (based on, say, GPT-4 speed on release). Then, by taking this inference penalty (with k= ~5), we'd be able to run ~1 million automated AI researchers at ~100x human speed.

[39] This source benchmarks throughput of Flash at ~6x GPT-4 Turbo, and GPT-4 Turbo was faster than original GPT-4. Latency is probably also roughly 10x faster.

instead? That would be 5+ OOMs in a year.

DON'T JUST IMAGINE 100 million junior software engineer interns here (we'll get those earlier, in the next couple years!). Real automated AI researchers be very smart—and in addition to their raw quantitative advantage, automated AI researchers will have other enormous advantages over human researchers:

- They'll be able to read every single ML paper ever written, have been able to deeply think about every single previous experiment ever run at the lab, learn in parallel from each of their copies, and rapidly accumulate the equivalent of millennia of experience. They'll be able to develop far deeper intuitions about ML than any human.

- They'll be easily able to write millions of lines of complex code, keep the entire codebase in context, and spend human-decades (or more) checking and rechecking every line of code for bugs and optimizations. They'll be superbly competent at all parts of the job.

- You won't have to individually train up each automated AI researcher (indeed, training and onboarding 100 million new human hires would be difficult). Instead, you can just teach and onboard one of them—and then make replicas. (And you won't have to worry about politicking, cultural acclimation, and so on, and they'll work with peak energy and focus day and night.)

- Vast numbers of automated AI researchers will be able to share context (perhaps even accessing each others' latent space and so on), enabling much more efficient collaboration and coordination compared to human researchers.

- And of course, however smart our initial automated AI researchers would be, we'd soon be able to make further OOM-jumps, producing even smarter models, even more capable at automated AI research.

Imagine an automated Alec Radford—*imagine 100 million au-*

tomated Alec Radfords.[40] I think just about every researcher at OpenAI would agree that if they had 10 Alec Radfords, let alone 100 or 1,000 or 1 million running at 10x or 100x human speed, they could very quickly solve very many of their problems. Even with various other bottlenecks (more in a moment), compressing a decade of algorithmic progress into a year as a result seems very plausible. (A 10x acceleration from a million times more research effort, which seems conservative if anything.)

That would be 5+ OOMs right there. 5 OOMs of algorithmic wins would be a similar scaleup to what produced the GPT-2-to-GPT-4 jump, a capability jump from ~a preschooler to ~a smart high schooler. Imagine such a qualitative jump *on top of* AGI, *on top of* Alec Radford.

It's strikingly plausible we'd go from AGI to superintelligence very quickly, perhaps in 1 year.

Possible bottlenecks

While this basic story is surprisingly strong—and is supported by thorough economic modeling work—there are some real and plausible bottlenecks that will probably slow down an automated-AI-research intelligence explosion.

I'll give a summary here, and then discuss these in more detail in the optional sections below for those interested:

- *Limited compute*: AI research doesn't just take good ideas, thinking, or math—but running experiments to get empirical signal on your ideas. A million times more research effort via automated research labor won't mean a million times faster progress, because compute will still be limited—and limited compute for experiments will be the bottleneck. Still, even if this won't be a 1,000,000x speedup, I find it hard to imagine that the automated AI researchers couldn't use the compute *at least* 10x more effectively: they'll be able to get incredible ML intuition (having internalized the whole ML literature and every previous experiment every run!) and

[40] Alec Radford is an incredibly gifted and prolific researcher/engineer at OpenAI, behind many of the most important advances, though he flies under the radar some.

centuries-equivalent of thinking-time to figure out exactly the right experiment to run, configure it optimally, and get maximum value of information; they'll be able to spend centuries-equivalent of engineer-time before running even tiny experiments to avoid bugs and get them right on the first try; they can make tradeoffs to economize on compute by focusing on the biggest wins; and they'll be able to try tons of smaller-scale experiments (and given effective compute scaleups by then, "smaller-scale" means being able to train 100,000 GPT-4-level models in a year to try architecture breakthroughs). Some human researchers and engineers are able to produce 10x the progress as others, even with the same amount of compute—and this should apply even moreso to automated AI researchers. I do think this is the most important bottleneck, and I address it in more depth below.

- *Complementarities/long tail:* A classic lesson from economics (cf Baumol's growth disease) is that if you can automate, say, 70% of something, you get some gains but quickly the remaining 30% become your bottleneck. For anything that falls short of full automation—say, really good copilots—human AI researchers would remain a major bottleneck, making the overall increase in the rate of algorithmic progress relatively small. Moreover, there's likely some long tail of capabilities required for automating AI research—the last 10% of the job of an AI researcher might be particularly hard to automate. This could soften takeoff some, though my best guess is that this only delays things by a couple years. Perhaps 2026/27-models speed are the proto-automated-researcher, it takes another year or two for some final unhobbling, a somewhat better model, inference speedups, and working out kinks to get to full automation, and finally by 2028 we get the 10x acceleration (and superintelligence by the end of the decade).

- *Inherent limits to algorithmic progress:* Maybe another 5 OOMs of algorithmic efficiency will be fundamentally impossible? I doubt it. While there will definitely be upper limits,[41] if we got 5 OOMs in the last decade, we should probably expect at least another decade's-worth of progress

[41] 25 OOMs of algorithmic progress on top of GPT-4, for example, are clearly impossible: that would imply it would be possible to train a GPT-4-level model with just a handful of FLOP.

to be possible. More directly, current architectures and train-
ing algorithms are still very rudimentary, and it seems that
much more efficient schemes should be possible. Biological
reference classes also support dramatically more efficient
algorithms being plausible.

- ***Ideas get harder to find, so the automated AI researchers
  will merely sustain, rather than accelerate, the current rate
  of progress:*** One objection is that although automated re-
  search would increase effective research effort a lot, ideas
  also get harder to find. That is, while it takes only a few
  hundred top researchers at a lab to sustain 0.5 OOMs/year
  today, as we exhaust the low-hanging fruit, it will take more
  and more effort to sustain that progress—and so the 100 mil-
  lion automated researchers will be merely what's necessary
  to sustain progress. I think this basic model is correct, but
  the empirics don't add up: the magnitude of the increase in
  research effort—a million-fold—is way, way larger than the
  historical trends of the growth in research effort that's been
  necessary to sustain progress. In econ modeling terms, it's a
  bizarre "knife-edge assumption" to assume that the increase
  in research effort from automation will be *just enough* to keep
  progress constant.

- ***Ideas get harder to find and there are diminishing returns, so
  the intelligence explosion will quickly fizzle:*** Related to the
  above objection, even if the automated AI researchers lead
  to an initial burst of progress, whether rapid progress can be
  sustained depends on the shape of the diminishing returns
  curve to algorithmic progress. Again, my best read of the
  empirical evidence is that the exponents shake out in favor
  of explosive/accelerating progress. In any case, the sheer
  size of the one-time boost—from 100s to 100s of millions of
  AI researchers—probably overcomes diminishing returns
  here for at least a good number of OOMs of algorithmic
  progress, even though it of course can't be *indefinitely* self-
  sustaining.

Overall, these factors may slow things down somewhat: the
most extreme versions of intelligence explosion (say, overnight)
seem implausible. And they may result in a somewhat longer

runup (perhaps we need to wait an extra year or two from more sluggish, proto-automated researchers to the true auto-mated Alec Radfords, before things kick off in full force). But they certainly don't rule out a very rapid intelligence explosion. A year—or at most just a few years, but perhaps even just a few months—in which we go from fully-automated AI researchers to vastly superhuman AI systems should be our mainline ex-pectation.

If you'd rather skip the in-depth discussions on the various bottle-necks below, click here to skip to the next section.

Limited compute for experiments (optional, in more depth)

The production function for algorithmic progress includes two complementary factors of production: research effort and experiment compute. The millions of automated AI researchers won't have any more compute to run their experiments on than human AI researchers; perhaps they'll just be sitting around waiting for their jobs to finish.

This is probably the most important bottleneck to the intelligence explosion. Ultimately this is a quantitative question—just how much of a bottleneck is it? On balance, I find it hard to believe that the 100 million Alec Radfords couldn't increase the marginal product of experiment com-pute by at least 10x (and thus, would still accelerate the pace of progress by 10x):

- *There's a lot you can do with smaller amounts of compute.* The way most AI research works is that you test things out at small scale—and then extrapolate via scaling laws. (Many key historical breakthroughs required only a very small amount of compute, e.g. the original Transformer was trained on just 8 GPUs for a few days.) And note that with ~5 OOMs of baseline scaleup in the next four years, "small scale" will mean GPT-4 scale—the auto-mated AI researchers will be able to run 100,000 GPT-4-level experiments on their training cluster in a year, and tens of millions of GPT-3-level experiments. (That's a lot

of potential-breakthrough new architectures they'll be able to test!)

– A lot of the compute goes into larger-scale validation of the final pretraining run—making sure you are getting a high-enough degree of confidence on marginal efficiency wins for your annual headline product—but if you're racing through the OOMs in the intelligence explosion, you could economize and just focus on the really big wins.

– As discussed in the previous piece, there are often *enormous* gains to be had from relatively low-compute "unhobbling" of models. These don't require big pretraining runs. It's highly plausible that that intelligence explosion starts off automated AI research e.g. discovering a way to do RL on top that gives us a couple OOMs via unhobbling wins (and then we're off to the races).

– *As the automated AI researchers find efficiencies, that'll let them run more experiments.* Recall the near-1000x cheaper inference in two years for equivalent-MATH performance, and the 10x general inference gains in the last year, discussed in the previous piece, from mere-human algorithmic progress. The first thing the automated AI researchers will do is quickly find similar gains, and in turn, that'll let them run 100x more experiments on e.g. new RL approaches. Or they'll be able to quickly make smaller models with similar performance in relevant domains (cf previous discussion of Gemini Flash, near-100x cheaper than GPT-4), which in turn will let them run many more experiments with these smaller models (again, imagine using these to try different RL schemes). There are probably other overhangs too, e.g. the automated AI researchers might be able to quickly develop much better distributed training schemes to utilize all the inference GPUs (probably at least 10x more compute right there). More generally, every OOM of training efficiency gains they find will give them an OOM

more of effective compute to run experiments on.

- *The automated AI researchers could be way more efficient.* It's hard to understate how many fewer experiments you would have to run if you just got it right on the first try—no gnarly bugs, being more selective about exactly what you are running, and so on. Imagine 1000 automated AI researchers spending a month-equivalent checking your code and getting the exact experiment right before you press go. I've asked some AI lab colleagues about this and they agreed: you should pretty easily be able to save 3x-10x of compute on most projects merely if you could avoid frivolous bugs, get things right on the first try, and only run high value-of-information experiments.

- *The automated AI researchers could have way better intuitions.*

  – Recently, I was speaking to an intern at a frontier lab; they said that their dominant experience over the past few months was suggesting many experiments they wanted to run, and their supervisor (a senior researcher) saying they could already predict the result beforehand so there was no need. The senior researcher's years of random experiments messing around with models had honed their intuitions about what ideas would work—or not. Similarly, it seems like our AI systems could easily get superhuman intuitions about ML experiments—they will have read the entire machine learning literature, be able to learn from every other experiment result and deeply think about it, they could easily be trained to predict the outcome of millions of ML experiments, and so on. And maybe one of the first things they do is build up a strong basic science of "predicting if this large scale experiment will be successful just after seeing the first 1% of training, or just after seeing the smaller scale version of this experiment", and so on.

  – Moreover, beyond really good intuitions about re-

search directions, as Jason Wei has noted, there are incredible returns to having great intuitions on the dozens of hyperparameters and details of an experiment,. Jason calls this ability to get things right on the first try based on intuition "yolo runs". (Jason says, "what I do know is that the people who can do this are surely 10-100x AI researchers.")

Compute bottlenecks will mean a million times more researchers won't translate into a million times faster research—thus not an overnight intelligence explosion. But the automated AI researchers will have extraordinary advantages over human researchers, and so it seems hard to imagine that they couldn't also find a way to use the compute at least 10x more efficiently/effectively—and so 10x the pace of algorithmic progress seems eminently plausible.

I'll take a moment here to acknowledge perhaps the most compelling formulation of the counterargument I've heard, by my friend James Bradbury: *if more ML research effort would so dramatically accelerate progress, why doesn't the current academic ML research community, numbering at least in the tens of thousands, contribute more to frontier lab progress?* (Currently, it seems like lab-internal teams, of perhaps a thousand in total across labs, shoulder most of the load for frontier algorithmic progress.) His argument is that the reason is that algorithmic progress is compute-bottlenecked: the academics just don't have enough compute.

Some responses:

- Quality-adjusted, I think academics are probably more in the thousands not tens of thousands (e.g., looking only at the top universities). This probably isn't substantially more than the labs combined. (And it's *way* less than the hundreds of millions of researchers we'd get from automated AI research.)

- Academics work on the wrong things. Up until very

recently (and perhaps still today?), the vast majority of the academic ML community wasn't even working on large language models. In terms of strong academics in academia working on large language models, it might be meaningfully fewer than researchers at labs combined?

- Even when the academics do work on things like LLM pretraining, they simply don't have access to the state-of-the-art—the large accumulated body of knowledge of tons of details on frontier model training inside labs. They don't know what problems are actually relevant, or can only contribute one-off results that nobody can really do anything with because their baselines were badly tuned (so nobody knows if their thing is actually an improvement).

- Academics are way worse than automated AI researchers: they can't work at 10x or 100x human speed, they can't read and internalize every ML paper ever written, they can't spend a decade checking every line of code, replicate themselves to avoid onboarding-bottlenecks, etc.

Another countervailing example to the academics argument: GDM is rumored to have way more experiment compute than OpenAI, and yet it doesn't seem like GDM is massively outpacing OpenAI in terms of algorithmic progress.

In general, I expect automated researchers will have a different style of research that plays to their strengths and aims to mitigate the compute bottleneck. *I think it's reasonable to be uncertain how this plays out, but it's unreasonable to be confident it won't be doable for the models to get around the compute bottleneck just because it'd be hard for humans to do so.*

- For example, they could just spend a lot of effort early on building up a basic science of "how to predict large scale results from smaller scale experiments". And I expect there's a lot that they could do that humans can't do, e.g. maybe things more like "predicting if this large scale experiment will be successful just after seeing the

first 1% of training". This seems pretty doable if you're a super strong automated researcher with very superhuman intuitions and this can save you a ton of compute.

- When I imagine AI systems automating AI research, I see them as compute-bottlenecked but making up for it in large part by thinking e.g. 1000x more (and faster) than humans would, and thinking at a higher level of quality than humans (e.g. because of the superhuman ML intuitions from being trained to predict the result of millions of experiments). Unless they're just much worse at thinking than engineering, I think this can make up for a lot, and this would be qualitatively different from academics.

(In addition to experiment compute, there's the additional bottleneck of eventually needing to run a big training run, something which currently takes months. But you can probably economize on those, doing only a handful during the year of intelligence explosion, taking bigger OOM leaps for each than labs currently do.[42] Or you could "spend" 1 out of the 5 OOMs of compute efficiency wins to do a training run in days rather than months.)

[42] Note that while I think this is likely, it's kind of scary: it means that rather than a fairly continuous series of big models, each somewhat better than the previous generation, downstream model intelligence might be more discrete/discontinuous. We might only do one or a couple of big runs during the intelligence explosion, banking multiple OOMs of algorithmic breakthroughs found at smaller scale for each.

Complementarities and long tails to 100% automation (optional, in more depth)

The classic economist objection to AI automation speeding up economic growth is that different tasks are complementary: so, for example, automating 80% of what labor humans did in 1800 didn't lead to a growth explosion or mass unemployment, but the remaining 20% became what all humans did and remained the bottleneck. (See e.g. a model of this here.).

I think the economists' *model* here is correct. But a key point is that I'm only talking about one currently-small part of the economy, rather than the economy as a whole. People may well still be getting haircuts normally during this time—robotics might not yet be worked out, AIs for every domain might not yet be worked out, the societal

rollout might not yet be worked out, etc.—but they *will* be able to do AI research. As discussed in the previous piece, I think the current course of AI progress is taking us to essentially drop-in remote workers as intelligent as the smartest humans; as discussed in this piece, the job of an AI researcher seems totally within the scope of what could be fully automated.

Still, in practice, I do expect somewhat of a long tail to get to truly 100% automation even for the job of an AI researcher/engineer; for example, we might first get systems that function *almost* as an engineer replacement, but still need some amount of human supervision.

In particular, I expect the level of AI capabilities to be somewhat uneven and peaky across domains: it might be a better coder than the best engineers while still having blindspots in some subset of tasks or skills; by the time it's human-level at whatever its worst at, it'll already be substantially superhuman at easier domains to train, like coding. (This is part of why I think they'll be able to use the compute more effectively than human researchers. By the time of 100% automation/the intelligence explosion starting, they'll already have huge advantages over humans in some domains. This will also have important implications for superalignment down the line, since it means that we'll have to align systems that are meaningfully superhuman in many domains in order to align even the first automated AI researchers.)

But I wouldn't expect that phase to last more than a few years; given the pace of AI progress, I think it would likely just be a matter of some additional "unhobbling" (removing some obvious limitation of the models that prevented it from doing the last mile) or another generation of models to get all the way.

Overall, this might soften takeoff some. Rather that 2027 AGI → 2028 Superintelligence, it might look more like:

- 2026/27: Proto-automated-engineer, but blind spots in

other areas. Speeds up work by 1.5x-2x already; progress begins gradually accelerating.

- 2027/28: Proto-automated-researchers, can automate >90%. Some remaining human bottlenecks, and hiccups in coordinating a giant organization of automated researchers to be worked out, but this already speeds up progress by 3x+. This quickly does the remaining necessary "unhobbling" takes us the remainder of the way to 100% automation.

- 2028/29: 10x+ pace of progress → superintelligence.

That's still very fast. . .

Fundamental limits to algorithmic progress (optional, in more depth)

There's probably a real cap on how much algorithmic progress is physically possible. (For example, 25 OOMs of algorithm progress seems impossible, since that would imply being able to train a GPT-4 level system in less than ~10 FLOPs.[43]) But something like 5 OOMs seems very much in the realm of possibilities; again, that would just require another decade of trend algorithmic efficiencies (not even counting algorithmic gains from unhobbling).

[43] Though you could get results that would take 25 more OOMs of hardware with current architecture!

Intuitively, it very much doesn't seem like we have exhausted all the low-hanging fruit yet, given how simple the biggest breakthroughs are—and how rudimentary and obviously hobbled current architectures and training techniques still seem to be. For example, I think it's pretty plausible that we'll bootstrap our way to AGI via AI systems that "think out loud" via chain-of-thought. But *surely* this isn't the most efficient way to do it, *surely* something that does this reasoning via internal states/recurrence/etc would be way more efficient. Or consider adaptive compute: Llama 3 still spends as much compute on predicting the "and" token as it does the answer to some complicated question, which seems clearly suboptimal. We're getting huge OOM algorithmic gains from even just small tweaks, while there are dozens of areas where much more effi-

cient architectures and training procedures could likely be found.

Biological references also suggest huge headroom. The human range of intelligence is very wide, for example, with only tiny tweaks to architecture. Humans have similar numbers of neurons as other animals, even though humans are much smarter than those animals. And current AI models are still many OOMs from the efficiency of a human brain; they can learn with a tiny fraction of the data (and thus tiny fraction of "compute") than AI models can, suggesting huge headroom for our algorithms and architecture.

Ideas get harder to find and diminishing returns (optional, in more depth)

As you pick the low-hanging fruit, ideas get harder to find. This is true in any domain of technological progress. Essentially, we see a straight line on a log-log curve: log(progress) is a function of log(cumulative research effort). Every OOM of further progress requires putting in more research effort than the last OOM.

This leads to two objections to the intelligence explosion:

1. Automated AI research will merely be what's necessary to sustain progress (rather than dramatically accelerating it).

2. A purely-algorithmic intelligence explosion would not be sustained / would quickly fizzle out as algorithmic progress gets harder to find / you hit diminishing marginal returns.

I spent a lot of time thinking about these sorts of models in a past life, when I was doing research in economics. (In particular, semi-endogenous growth theory is the standard model of technological progress, capturing these two competing dynamics of growing research effort and ideas getting harder to find.)

In short, I think the underlying model behind these objections is sound, but how it shakes out is an empirical question—and I think they get the empirics wrong.

The key question is essentially: for every 10x of progress, does further progress become *more* or *less* than 10x harder? Napkin math (along the lines of how this is done in the economic literature) helps us bound this.

- Suppose we take the ~0.5 OOMs/year trend rate of algorithmic progress seriously; that implies a 100x of progress in 4 years.

- However, quality-adjusted headcount / research effort at a given leading AI lab has definitely grown by <100x in 4 years. *Maybe* it's increased 10x (from 10s to 100s of people working on relevant stuff at a given lab), but even that is unclear quality-adjusted.

- And yet, algorithmic progress seems to be sustained.

Thus, in response to objection 1, we can note that the ~million-fold increase in research effort will simply be a much larger increase than what would merely be necessary to sustain progress. Maybe we'd need on the order of thousands of researchers working on relevant research at a lab in 4 years to sustain progress; the 100 million Alec Radfords would still be an *enormous* increase, and surely lead to massive acceleration. It's just a bizarre "knife-edge" assumption to think that automated research would be *just enough* to sustain the existing pace of progress. (And that's not even counting thinking at 10x human speed and all the other advantages the AI systems will have over human researchers.)

In response to objection 2, we can note two things:

- First, the mathematical condition noted above. Given that, based on our napkin math, quality-adjusted research effort needed to grow «100x while we did 100x of algorithmic progress, it pretty strongly seems that the

shape of the returns curve shakes out in favor of self-sustaining progress.[44]

- Secondly, the returns curve doesn't even *need* to shake out in favor of a fully sustained chain reaction for us to get a bounded-but-many-OOM surge of algorithmic progress. Essentially, it doesn't need to be a "growth effect"; a large enough "level effect" would be enough. That is, a million-fold increase in research effort (combined with the many other advantages automated researchers would have over human researchers) would be such a large one-time boost that even if the chain reaction isn't fully self-sustaining, it could lead to a very sizeable (many OOMs) one-time gain.[45]

On net, while obviously it won't be unbounded and I have a lot of uncertainty over just how far it'll go, I think something like a 5 OOM intelligence explosion purely from algorithmic gains / automated AI research seems highly plausible.

Tom Davidson and Carl Shulman have also looked at the empirics of this in a growth-modeling framework and come to similar conclusions. Epoch AI has also done some recent work on the empirics, also coming to the conclusion that empirical returns to algorithmic R&D favors explosive growth, with a helpful writeup of the implications.

The power of superintelligence

WHETHER OR NOT you agree with the strongest form of these arguments—whether we get a <1 year intelligence explosion, or it takes a few years—it is clear: we must confront the possibility of superintelligence.

The AI systems we'll likely have *by the end of this decade* will be unimaginably powerful.

- Of course, they'll be *quantitatively* superhuman. On our fleets

[44] 100x progress → 100x more automated research effort, but you, say, only needed 10x more research effort to keep it going and do the next 100x, so the returns are good enough for explosive progress to be sustained.

[45] Analogously, in semi-endogenous economic growth theory, boosting science investment from 1% to 20% of GDP won't make growth rates higher forever—eventually diminishing returns would lead things to return to the old growth rate—but the "level effect" it would lead to would be so large as to dramatically speed up growth for decades.

of 100s of millions of GPUs by the end of the decade, we'll be able to run a civilization of billions of them, and they will be able to "think" orders of magnitude faster than humans. They'll be able to quickly master any domain, write trillions of lines of code, read every research paper in every scientific field ever written (they'll be perfectly interdisciplinary!) and write new ones before you've gotten past the abstract of one, learn from the parallel experience of every one of its of copies, gain billions of human-equivalent years of experience with some new innovation in a matter of weeks, work 100% of the time with peak energy and focus and won't be slowed down by that one teammate who is lagging, and so on.

• More importantly—but harder to imagine—they'll be *qualitatively* superhuman. As a narrow example of this, large-scale RL runs have been able to produce completely novel and creative behaviors beyond human understanding, such as the famous move 37 in AlphaGo vs. Lee Sedol. Superintelligence will be like this across many domains. It'll find exploits in human code too subtle for any human to notice, and it'll generate code too complicated for any human to understand even if the model spent decades trying to explain it. Extremely difficult scientific and technological problems that a human would be stuck on for decades will seem just *so obvious* to them. We'll be like high-schoolers stuck on Newtonian physics while it's off exploring quantum mechanics.

As an example of how *wild* this could be, look at some Youtube videos of video game speedruns, such as this one of beating Minecraft in 20 seconds. (If you have no idea what's going on in this video, you're in good company; even most normal players of Minecraft have almost no clue what's going on.)

Now imagine this applied to all domains of science, technology, and the economy. The error bars here, of course, are extremely large. Still, *this is happening,* and it's important to consider just how consequential this would be.
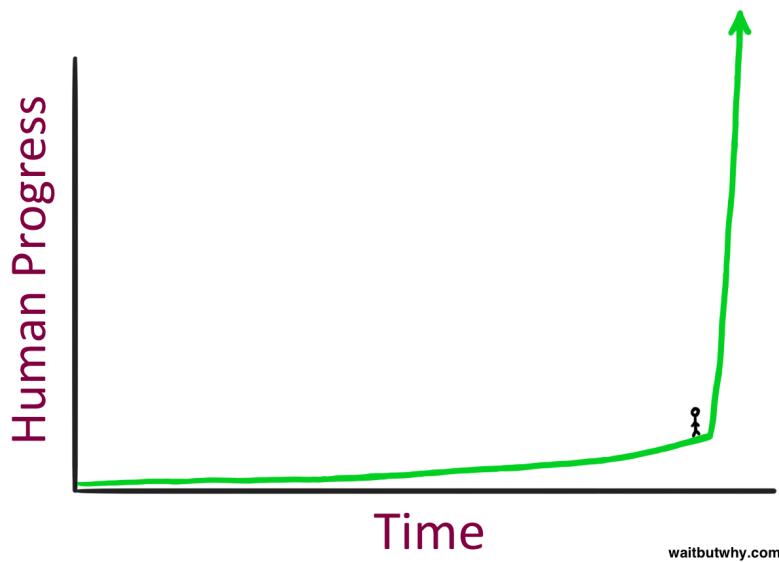
Figure 25: What does it feel like to stand here? Illustration from Wait But Why/Tim Urban.

In the intelligence explosion, explosive progress was initially only in the narrow domain of automated AI research. As we get superintelligence, and apply our billions of (now superintelligent) agents to R&D across many fields, I expect explosive progress to broaden:

- *An AI capabilities explosion.* Perhaps our initial AGIs had limitations that prevented them fully automating work in some other domains (rather than just in the AI research domain); automated AI research will quickly solve these, enabling automation of any and all cognitive work.

- *Solve robotics.* Superintelligence won't stay purely cognitive for long. Getting robotics to work well is primarily an ML algorithms problem (rather than a hardware problem), and our automated AI researchers will likely be able to solve it (more below!). Factories would go from human-run, to AI-directed using human physical labor, to soon being fully run by swarms of robots.

- *Dramatically accelerate scientific and technological progress.* Yes, Einstein alone couldn't develop neuroscience and build a semiconductor industry, but a billion superintelligent automated scientists, engineers, technologists, and robot

technicians (with the robots moving at 10x or more human speed!)[46] would make extraordinary advances in many fields in the space of years. (Here's a nice short story visualizing what AI-driven R&D might look like.) The billion superintelligences would be able to compress the R&D effort humans researchers would have done in the next century into years. Imagine if the technological progress of the 20th century were compressed into less than a decade. We would have gone from flying being thought a mirage, to airplanes, to a man on the moon and ICBMs in a matter of years. *This is what I expect the 2030s to look like across science and technology.*

- *An industrial and economic explosion.* Extremely accelerated technological progress, combined with the ability to automate all human labor, could dramatically accelerate economic growth (think: self-replicating robot factories quickly covering all of the Nevada desert[47]). The increase in growth probably wouldn't just be from 2%/year to 2.5%/year; rather, this would be a fundamental shift in the growth regime, more comparable to the historical step-change from very slow growth to a couple percent a year with the industrial revolution. We could see economic growth rates of 30%/year and beyond, quite possibly multiple doublings a year. This follows fairly straightforwardly from economists' models of economic growth. To be sure, this may well be delayed by societal frictions; arcane regulation might ensure lawyers and doctors still need to be human, even if AI systems were much better at those jobs; surely sand will be thrown into the gears of rapidly expanding robo-factories as society resists the pace of change; and perhaps we'll want to retain human nannies; all of which would slow the growth of the overall GDP statistics. Still, in whatever domains we remove human-created barriers (e.g., competition might force us to do so for military production), we'd see an industrial explosion.

[46] The 10x speed robots doing physical R&D in the real world is the "slow version"; in reality the superintelligences will try to do as much R&D as possible in simulation, like AlphaFold or manufacturing "digital twins".

[47] Why isn't "factorio-world"—build a factory, that produces more factories, producing even more factories, doubling factories until eventually your entire planet is quickly covered in factories—possible today? Well, labor is constrained—you can accumulate capital (factories, tools, etc.), but that runs into diminishing returns because it's constrained by a fixed labor force. With robots and AI systems being able to fully automate labor, that removes that constraint; robo-factories could produce more robo-factories in an ~unconstrained way, leading to an industrial explosion. See more economic growth models of this here.

Growth mode	Date began to dominate	Doubling time of global economy (years)
Hunting	2,000,000 B.C.	230,000
Farming	4700 B.C.	860
Science/commerce	1730 A.D.	58
Industry	1903 A.D.	15
Superintelligence?	2030 A.D.?	???

Table 3: A shift in the growth regime is not unprecedented: as civilization went from hunting, to farming, to the blossoming of science and commerce, to industry, the pace of global economic growth accelerated. Superintelligence could kick off another shift in growth mode. Based on Robin Hanson's "Long-run growth as a sequence of exponential modes".

- *Provide a decisive and overwhelming military advantage.* Even early cognitive superintelligence might be enough here; perhaps some superhuman hacking scheme can deactivate adversary militaries. In any case, military power and technological progress has been tightly linked historically, and with extraordinarily rapid technological progress will come concomitant military revolutions. The drone swarms and roboarmies will be a big deal, but they are just the beginning; we should expect completely new kinds of weapons, from novel WMDs to invulnerable laser-based missile defense to things we can't yet fathom. Compared to pre-superintelligence arsenals, it'll be like 21st century militaries fighting a 19th century brigade of horses and bayonets. (I discuss how superintelligence could lead to a decisive military advantage in a later piece.)

- *Be able to overthrow the US government.* Whoever controls superintelligence will quite possibly have enough power to seize control from pre-superintelligence forces. Even without robots, the small civilization of superintelligences would be able to hack any undefended military, election, television, etc. system, cunningly persuade generals and electorates, economically outcompete nation-states, design new synthetic bioweapons and then pay a human in bitcoin to synthesize it, and so on. In the early 1500s, Cortes and about 500 Spaniards conquered the Aztec empire of several million; Pizarro and ~300 Spaniards conquered the Inca empire of several million; Alfonso and ~1000 Portuguese conquered the Indian Ocean. They didn't have god-like power, but the Old World's technological edge and an advantage in strate-

gic and diplomatic cunning led to an utterly decisive advantage. Superintelligence might look similar.

SITUATIONAL AWARENESS | Leopold Aschenbrenner

Figure 26: Explosive growth starts in the narrower domain of AI R&D; as we apply superintelligence to R&D in other fields, explosive growth will broaden.
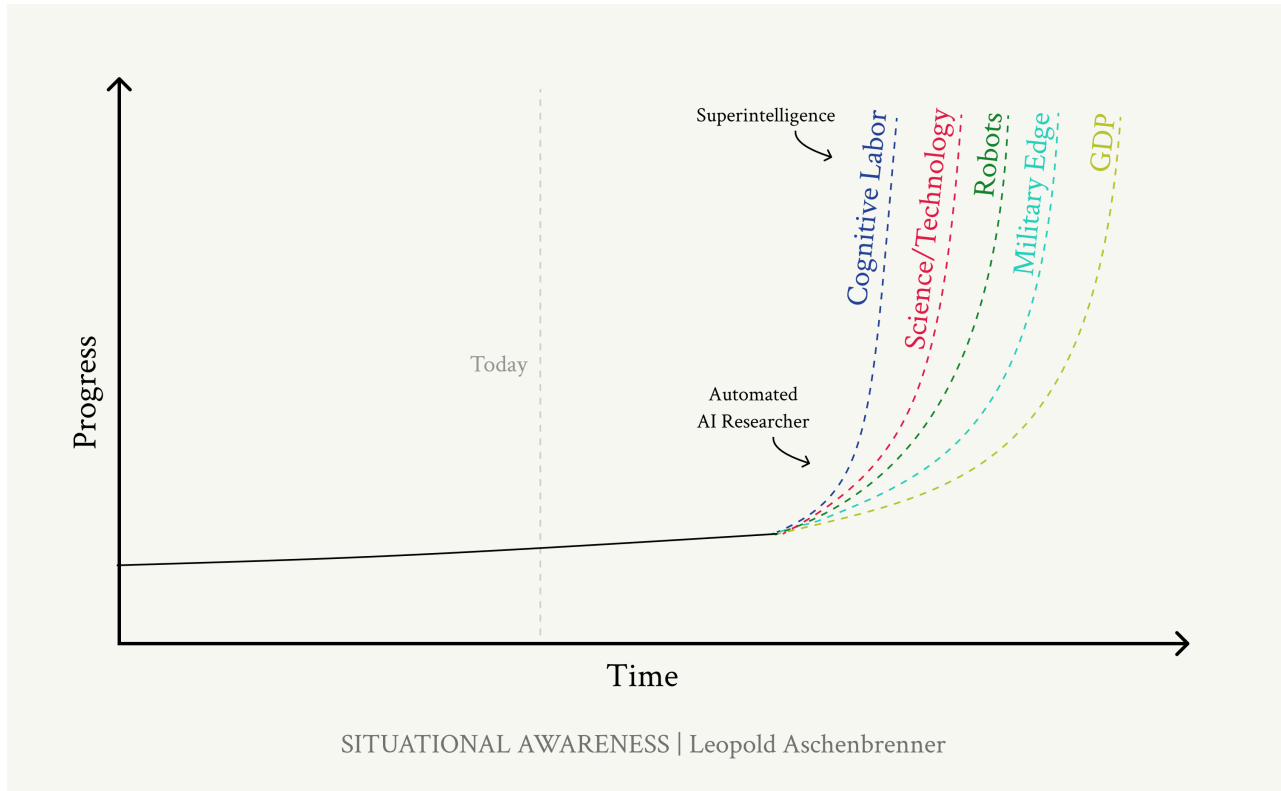
Robots. A common objection to claims like those here is that, even if AI can do cognitive tasks, robotics is lagging way behind and so will be a brake on any real-world impacts.

I used to be sympathetic to this, but I've become convinced robots will not be a barrier. For years people claimed robots were a hardware problem—but robot hardware is well on its way to being solved.

Increasingly, it's clear that robots are an *ML algorithms* problem. LLMs had a much easier way to bootstrap: you had an entire internet to pretrain on. There's no similarly large dataset for robot actions, and so it requires more nifty approaches (e.g. using multimodal models as a base, then using synthetic data/simulation/clever RL) to train them.

There's a ton of energy directed at solving this now. But even if we don't solve it before AGI, our hundreds of millions of AGIs/superintelligences will make *amazing* AI researchers (as is the central argument of this piece!), and it seems very likely that they'll figure out the ML to make amazing robots work.

As such, while it's plausible that robots might cause a few years of delay (solving the ML problems, testing in the physical world in a way that is fundamentally slower than testing in simulation, ramping up initial robot production before the robots can build factories themselves, etc.)—I don't think it'll be more than that.

How ALL OF THIS plays out over the 2030s is hard to predict (and a story for another time). But one thing, at least, is clear: we will be rapidly plunged into the most extreme situation humanity has ever faced.

Human-level AI systems, AGI, would be highly consequential in their own right—but in some sense, they would simply be a more efficient version of what we already know. But, very plausibly, within just a year, we would transition to much more alien systems, systems whose understanding and abilities— whose raw power—would exceed those even of humanity combined. There is a real possibility that we will lose control, as we are forced to hand off trust to AI systems during this rapid transition.

More generally, everything will just start happening *incredibly fast*. And the world will start going insane. Suppose we had gone through the geopolitical fever-pitches and man-made perils of the 20th century in mere years; that is the sort of situation we should expect post-superintelligence. By the end of it, superintelligent AI systems will be running our military and economy. During all of this insanity, we'd have extremely scarce time to make the right decisions. The challenges will be immense. It will take everything we've got to make it through in one piece.

The intelligence explosion and the immediate post-superintelligence period will be one of the most volatile, tense, dangerous, and wildest periods ever in human history.

And by the end of the decade, we'll likely be in the midst of it.

CONFRONTING THE POSSIBILITY of an intelligence explosion—the emergence of superintelligence—often echoes the early debates around the possibility of a nuclear chain reaction—and the atomic bomb it would enable. HG Wells predicted the atomic bomb in a 1914 novel. When Szilard first conceived of the idea of a chain reaction in 1933, he couldn't convince anyone of it; it was pure theory. Once fission was empirically discovered in 1938, Szilard freaked out again and argued strongly for secrecy, and a few people started to wake up to the possibility of a bomb. Einstein hadn't considered the possibility of a chain reaction, but when Szilard confronted him, he was quick to see the implications and willing to do anything that was needed to be done; he was willing to sound the alarm, and wasn't afraid of sounding foolish. But Fermi, Bohr, and most scientists thought the "conservative" thing was to play it down, rather than take seriously the extraordinary implications of the possibility of a bomb. Secrecy (to avoid sharing their advances with the Germans) and other all-out efforts seemed absurd to them. A chain reaction sounded too crazy. (Even when, as it turned out, a bomb was but half a decade from becoming reality.)

We must once again confront the possibility of a chain reaction. Perhaps it sounds speculative to you. But among senior scientists at AI labs, many see a rapid intelligence explosion as strikingly plausible. They can *see* it. Superintelligence is possible.

III. The Challenges

IIIa. Racing to the Trillion-Dollar Cluster

The most extraordinary techno-capital acceleration has been set in motion. As AI revenue grows rapidly, many trillions of dollars will go into GPU, datacenter, and power buildout before the end of the decade. The industrial mobilization, including growing US electricity production by 10s of percent, will be intense.

> You see, I told you it couldn't be done without turning the whole country into a factory. You have done just that.
>
> NIELS BOHR
> (to Edward Teller, upon learning of the scale of the
> Manhattan Project in 1944)

THE RACE TO AGI won't just play out in code and behind laptops—it'll be a race to mobilize America's industrial might. Unlike anything else we've recently seen come out of Silicon Valley, AI is a massive industrial process: each new model requires a giant new cluster, soon giant new power plants, and eventually giant new chip fabs. The investments involved are staggering. But behind the scenes, they are already in motion.

In this chapter, I'll walk you through numbers to give you a sense of what this will mean:

- As revenue from AI products grows rapidly—plausibly hitting a $100B annual run rate for companies like Google or Microsoft by ~2026, with powerful but pre-AGI systems—

Figure 27: The trillion-dollar cluster.
Credit: DALLE.

that will motivate ever-greater capital mobilization, and total AI investment could be north of $1T *annually* by 2027.

- We're on the path to individual training clusters costing $100s of billions by 2028—clusters requiring power equivalent to a small/medium US state and more expensive than the International Space Station.

- By the end of the decade, we are headed to $1T+ individual training clusters, requiring power equivalent to >20% of US electricity production. Trillions of dollars of capex will churn out 100s of millions of GPUs per year overall.

Nvidia shocked the world as its datacenter sales exploded from about $14B annualized to about $90B annualized in the last year. But that's still just the very beginning.

Training compute

Earlier, we found a roughly ~0.5 OOMs[48]/year trend growth of AI training compute. If this trend were to continue for the rest of the decade, what would that mean for the largest training clusters?

Year	OOMs	H100s-equivalent	Cost	Power	Power reference class
2022	~GPT-4 cluster	~10k	~$500M	~10 MW	~10,000 average homes
~2024	+1 OOM	~100k	$billions	~100MW	~100,000 homes
~2026	+2 OOMs	~1M	$10s of billions	~1 GW	The Hoover Dam, or a large nuclear reactor
~2028	+3 OOMs	~10M	$100s of billions	~10 GW	A small/medium US state
~2030	+4 OOMs	~100M	$1T+	~100GW	>20% of US electricity production

Table 4: Scaling the largest training clusters, rough back-of-the-envelope calculations. For details on the calculations, see Appendix.

This may seem hard to believe—but it appears to be happening. Zuck bought 350k H100s. Amazon bought a 1GW datacenter campus next to a nuclear power plant. Rumors suggest a 1GW, 1.4M H100-equivalent cluster (~2026-cluster) is being built in Kuwait. Media report that Microsoft and OpenAI are rumored to be working on a $100B cluster, slated for 2028 (a cost comparable to the International Space Station!). And as each generation of models shocks the world, further acceleration may yet be in store.

Perhaps the wildest part is that willingness-to-spend doesn't even seem to be the binding constraint at the moment, at least for training clusters. It's finding the infrastructure itself: "Where do I find 10GW?" (power for the $100B+, trend 2028 cluster) is a favorite topic of conversation in SF. What any compute guy is thinking about is securing power, land, permitting, and datacenter construction.[49] While it may take you a year of waiting to get the GPUs, the lead times for these are much longer still.

The trillion-dollar cluster—+4 OOMs from the GPT-4 cluster, the ~2030 training cluster on the current trend—will be a truly extraordinary effort. The 100GW of power it'll require is equivalent to >20% of US electricity production; imagine not just a simple warehouse with GPUs, but hundreds of power plants. Perhaps it will take a national consortium.

(Note that I think it's pretty likely we'll only need a ~$100B cluster, or less, for AGI. The $1T cluster might be what we'll train and run superintelligence on, or what we'll use for AGI if AGI is harder than expected. In any case, in a post-AGI world, having the most compute will probably still really matter.)

Overall compute

The above are just rough numbers for the largest training clusters. Overall investment is likely to be much larger still: a large fraction of GPUs will probably be be used for inference[50] (GPUs to actually run the AI systems for products), and there could be multiple players with giant clusters in the race.

My rough estimate is that 2024 will already feature $100B-$200B of AI investment:

- Nvidia datacenter revenue will hit a ~$25B/quarter run rate soon, i.e. ~$100B of capex flowing via Nvidia alone. But of course, Nvidia isn't the only player (Google's TPUs are great too!), and close to half of datacenter capex is on things other than the chips (site, building, cooling, power, etc.).[51]

- Big tech has been dramatically ramping their capex numbers: Microsoft and Google will likely do $50B+[52], AWS and Meta $40B+, in capex this year. Not all of this is AI, but combined their capex will have grown $50B-$100B year-over-year because of the AI boom, and even then they are still cutting back on other capex to shift even more spending to AI. Moreover, other cloud providers, companies (e.g., Tesla is spending $10B on AI this year), and nation-states are investing in AI as well.

[50] See, for example, Zuck here; only ~45k of his H100s are in his largest training clusters, the vast majority of his 350k H100s for inference. Meta likely has heavier inference needs than other players, who serve fewer customers so far, but as everyone else's AI products scale, I expect inference to become a strong majority of the GPUs.

[51] For example, this total-cost-of-ownership analysis estimates that around 40% of a large cluster cost is the H100 GPUs itself, and another 13% goes to Nvidia for Infiniband networking. That said, excluding cost of capital in that calculation would mean the GPUs are about 50% of the cost, and with networking Nvidia gets a bit over 60% of the cost of the cluster.

[52] And apparently, despite Microsoft growing capex by 79% compared to a year ago in a recent quarter, their AI cloud demand still exceeds supply!

Quarterly NVIDIA Datacenter Revenue
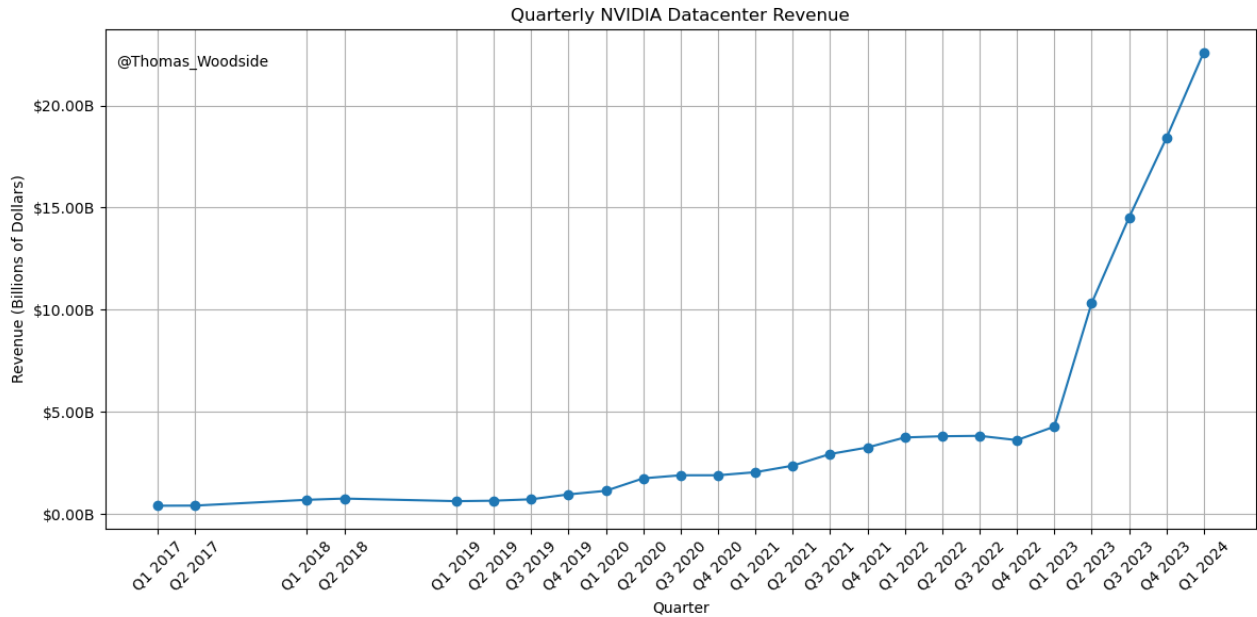
@Thomas_Woodside

Figure 28: Quarterly Nvidia datacenter revenue. Plot by Thomas Woodside

MSFT, GOOG, META Capital Expenditures
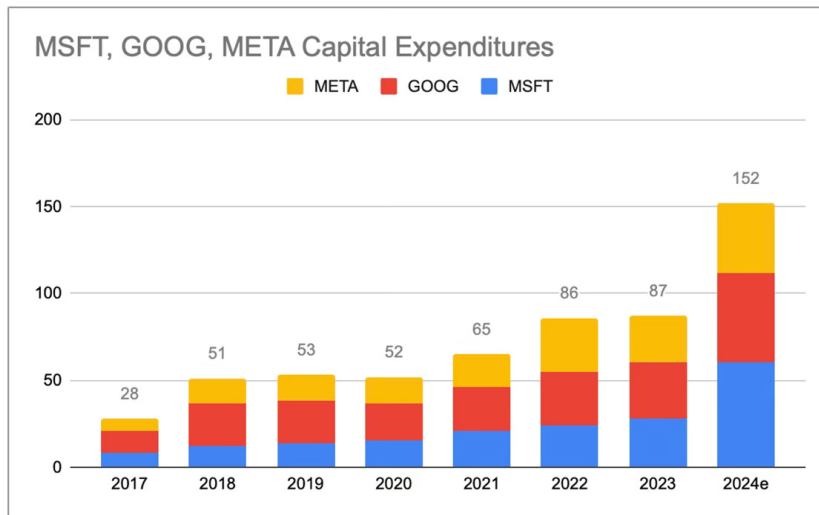
Figure 29: Big tech capex is growing extremely rapidly since ChatGPT unleashed the AI boom. Source.

Let's play this forward. My best guess is overall compute investments will grow more slowly than the 3x/year largest training clusters, let's say 2x/year.[53]

And these aren't just my idiosyncratic numbers. AMD forecasted a $400B AI accelerator market by 2027, implying $700B+ of total AI spending, pretty close to my numbers (and they are surely much less "AGI-pilled" than I am). Sam Altman is re-

[53] A larger fraction of global GPU production will probably be going to the largest training cluster in the future than today, e.g. because of a consolidation to just a few leading labs, rather than many companies having frontier-model-scale clusters. Regarding power: Of course, not all of this will be in the US, but to give a reference class.

Year	Annual investment	AI accelerator shipments (in H100s-equivalent)	Power as % of US electricity production	Chips as % of current leading-edge TSMC wafer production
2024	~$150B	~5-10M	1-2%	5-10%
~2026	~$500B	~10s of millions	5%	~25%
~2028	~$2T	~100M	20%	~100%
~2030	~$8T	~100s of millions	100%	4x current capacity

Table 5: Playing forward trends on total world AI investment. Rough back-of-the-envelope calculation. For some further details on the calculations, see Appendix.

ported to be in talks to raise funds for a project of "up to $7T" in capex to build out AI compute capacity (the number was widely mocked, but it seems less crazy if you run the numbers here...). One way or another, this massive scaleup is happening.

Will it be done? Can it be done?

THE SCALE OF INVESTMENT postulated here may seem fantastical. But both the demand-side and the supply-side seem like they could support the above trajectory. The economic returns justify the investment, the scale of expenditures is not unprecedented for a new general-purpose technology, and the industrial mobilization for power and chips is doable.

AI revenue

Companies will make large AI investments if they expect the economic returns to justify it.

Reports suggest OpenAI was at a $1B revenue run rate in August 2023, and a $2B revenue run rate in February 2024. That's roughly a doubling every 6 months. If that trend holds, we should see a ~$10B annual run rate by late 2024/early 2025, even without pricing in a massive surge from any next-generation model. One estimate puts Microsoft at ~$5B of incremental AI revenue already.

So far, every 10x scaleup in AI investment seems to yield the necessary returns. GPT-3.5 unleashed the ChatGPT mania. The estimated $500M cost for the GPT-4 cluster would have been paid off by the reported billions of annual revenue for Microsoft and OpenAI (see above calculations), and a "2024-class" training cluster in the billions will easily pay off if Microsoft/OpenAI AI revenue continues on track to a $10B+ revenue run rate. The boom is investment-led: it takes time from a huge order of GPUs to build the clusters, build the models, and roll them out, and the clusters being planned today are many years out. But if the returns on the last GPU order keep materializing, investment will continue to skyrocket (and outpace revenue), plowing in even more capital in a bet that the next 10x will keep paying off.

A key milestone for AI revenue that I like to think about is: when will a big tech company (Google, Microsoft, Meta, etc.) hit a $100B revenue run rate from AI (products and API)? These companies have on the order of $100B-$300B of revenue today; $100B would thus start representing a very substantial fraction of their business. Very naively extrapolating out the doubling every 6 months, supposing we hit a $10B revenue run rate in early 2025, suggests this would happen mid-2026.

That may seem like a stretch, but it seems to me to require surprisingly little imagination to reach that milestone. For example, there are around 350 million paid subscribers to Microsoft Office—could you get a third of these to be willing to pay $100/month for an AI add-on? For an average worker, that's only a few hours a month of productivity gained; models powerful enough to make that justifiable seem very doable in the next couple years.[54]

It's hard to understate the ensuing reverberations. This would make AI products the biggest revenue driver for America's largest corporations, and *by far* their biggest area of growth. Forecasts of overall revenue growth for these companies would skyrocket. Stock markets would follow; we might see our first $10T company soon thereafter. Big tech at this point would be willing to go all out, each investing many hundreds of billions (at least) into further AI scaleout. We probably see our first

[54] A big uncertainty for me is what the lags are for the technology to diffuse and be adopted. I think it's plausible revenue is slowed because intermediate, pre-AGI models take a lot of "schlep" to properly integrate into company workflows; historically, it's taken a while to fully harvest the productivity gains from new general purpose technologies. This is where the "sonic boom" discussion earlier comes in: as we "unhobble" models and they start looking more like agents/drop-in remote workers,, deploying them becomes much easier. Rather than having to completely remake some workflow to harvest a 25% productivity gain from a GPT-chatbot, instead you'll get models that you can onboard and work with as you would a new coworker (e.g., just directly substitute for an engineer, rather than needing to train up engineers to use some new tool). Or, in the extreme, and later on: you won't need to completely redesign a factory to work with some new tool, you'll just bring in the humanoid robots.

That said, this may lead to some discontinuity in economic value and revenue generated, depending on how quickly we can "unhobble" models.

many-hundred-billion dollar corporate bond sale then.[55]

[55] What will happen to interest rates will be interesting... see Tyler Cowen here; Chow, Mazlish and Halperin here.

Beyond $100B, it gets harder to see the contours. But if we are truly on the path to AGI, the returns will be there. White-collar workers are paid tens of trillions of dollars in wages annually worldwide; a drop-in remote worker that automates even a fraction of white-collar/cognitive jobs (imagine, say, a truly automated AI coder) would pay for the trillion-dollar cluster. If nothing else, the national security import could well motivate a government project, bundling the nation's resources in the race to AGI (more later).

Historical precedents

$1T/year of total annual AI investment by 2027 seems outrageous. But it's worth taking a look at other historical reference classes:

- In their peak years of funding, the Manhattan and Apollo programs reached 0.4% of GDP, or ~$100 billion annually today (surprisingly small!). At $1T/year, AI investment would be about 3% of GDP.

- Between 1996–2001, telecoms invested nearly $1 trillion in today's dollars in building out internet infrastructure.

- From 1841 to 1850, private British railway investments totaled a cumulative ~40% of British GDP at the time. A similar fraction of US GDP would be equivalent to ~$11T over a decade.

- Many trillions are being spent on the green transition.

- Rapidly-growing economies often spend a high fraction of their GDP on investment; for example, China has spent more than 40% of its GDP on investment for two decades (equivalent to $11T *annually* given US GDP).

- In the historically most exigent national security circumstances—wartime—borrowing to finance the national effort has often comprised enormous fractions of GDP. During WWI, the

UK and France, and Germany borrowed over 100% of their GDPs while the US borrowed over 20%; during WWII, the UK and Japan borrowed over 100% of their GDPs while the US borrowed over 60% of GDP (equivalent to over $17T today).

$1T/year of total AI investment by 2027 would be dramatic—among the very largest capital buildouts ever—but would not be unprecedented. And a trillion-dollar individual training cluster by the end of the decade seems on the table.[56]

Power

Probably the single biggest constraint on the supply-side will be power. Already, at nearer-term scales (1GW/2026 and especially 10GW/2028), power has become the binding constraint: there simply isn't much spare capacity, and power contracts are usually long-term locked-in. And building, say, a new gigawatt-class nuclear power plant takes a decade. (I'll wonder when we'll start seeing things like tech companies buying aluminum smelting companies for their gigawatt-class power contracts.[57])

[56] And, farther out, but if AGI truly led to substantial increases in economic growth, $10T+ annually would start being plausible—the reference class being investment rates of countries during high-growth periods.

[57] "Since 2011, the Alouette Smelter uses 930 MW electricity at maximum production capacity."

American Power Generation

Total Electricity Generation

Total AI Demand*

Largest Training Cluster*

*Back-of-the-envelope calculations.

SITUATIONAL AWARENESS | Leopold Aschenbrenner

Figure 30: Comparing trends on total US electricity production to our back-of-the-envelope estimates on AI electricity demands.

Total US electricity generation has barely grown 5% in the last

decade. Utilities are starting to get excited about AI (instead of 2.6% growth over the next 5 years, they now estimate 4.7%!). But they're barely pricing in what's coming. The trillion-dollar, 100GW cluster alone would require ~20% of current US electricity generation in 6 years; together with large inference capacity, demand will be multiples higher.

To most, this seems completely out of the question. Some are betting on Middle Eastern autocracies, who have been going around offering boundless power and giant clusters to get their rulers a seat at the AGI-table.

But it's totally possible to do this in the United States: we have abundant natural gas.[58]

- Powering a 10GW cluster would take only a few percent of US natural gas production and could be done rapidly.

- Even the 100GW cluster is surprisingly doable.

  - Right now the Marcellus/Utica shale (around Pennsylvania) alone is producing around 36 billion cubic feet a day of gas; that would be enough to generate just under 150GW continuously with generators (and combined cycle power plants could output 250 GW due to their higher efficiency).

  - It would take about ~1200 new wells for the 100GW cluster.[59] Each rig can drill roughly 3 wells per month, so 40 rigs (the current rig count in the Marcellus) could build up the production base for 100GW in less than a year.[60] The Marcellus had a rig count of ~80 as recently as 2019 so it would not be taxing to add 40 rigs to build up the production base.[61]

  - More generally, US natural gas production has more than doubled in a decade; simply continuing that trend could power multiple trillion-dollar datacenters.[62]

  - The harder part would be building enough generators / turbines; this wouldn't be trivial, but it seems doable with

[58] Thanks to Austin Vernon (private correspondence) for helping with these estimates.

[59] New wells produce around 0.01 BCF per day.

[60] Each well produces ~20 BCF over its lifetime, meaning two new wells a month would replace the depleted reserves. That is, it would need only one rig to maintain the production.

[61] Though it would be more efficient to add less rigs and build up over a longer time frame than 10 months.

[62] A cubic foot of natural gas generates about 0.13 kWh. Shale gas production was about ~70 billion cubic feet per day in the US in 2020. Suppose we doubled production again, and the extra capacity all went to compute clusters. That's 3322 TWh/year of electricity, or enough for almost 4 100GW clusters.

about \$100B of capex[63] for 100GW of natural gas power plants. Combined cycle plants can be built in about two years; the timeline for generators would be even shorter still.[64]

The barriers to even trillions of dollars of datacenter buildout in the US are entirely self-made. Well-intentioned but rigid climate commitments (not just by the government, but green datacenter commitments by Microsoft, Google, Amazon, and so on) stand in the way of the obvious, fast solution. At the very least, even if we won't do natural gas, a broad deregulatory agenda would unlock the solar/batteries/SMR/geothermal megaprojects. Permitting, utility regulation, FERC regulation of transmission lines, and NEPA environmental review makes things that should take a few years take a decade or more. We don't have that kind of time.

We're going to drive the AGI datacenters to the Middle East, under the thumb of brutal, capricious autocrats. I'd prefer clean energy too—but this is simply too important for US national security. We will need a new level of determination to make this happen. The power constraint can, must, and will be solved.

Chips

While chips are usually what comes to mind when people think about AI-supply-constraints, they're likely a smaller constraint than power. Global production of AI chips is still a pretty small percent of TSMC-leading-edge production, likely less than 10%. There's a lot of room to grow via AI becoming a larger share of TSMC production.

Indeed, 2024 production of AI chips (~5-10M H100-equivalents) would already be almost enough for the \$100s of billion cluster (if they were all diverted to one cluster). From a pure logic fab standpoint ~100% of TSMC's output for a year could already support the trillion-dollar cluster (again if all the chips went to one datacenter).[65] Of course, not all of TSMC will be able to be diverted to AI, and not all of AI chip production for a year will be for one training cluster. Total AI chip demand (including

[63] The capex costs for natural gas power plants seem to be under \$1000 per kW, meaning the capex for 100GW of natural gas power plants would be about \$100 billion.

[64] Solar and batteries aren't a totally crazy alternative, but it does just seem rougher than natural gas. I did appreciate Casey Handmer's calculation of tiling the Earth in solar panels:
"With current GPUs, the global solar datacenter's compute is equivalent to ~150 billion humans, though if our computers can eventually match [human brain] efficiency, we could support more like 5 quadrillion AI souls."

[65] This poses the interesting question of why power requirements are going up so much before chip fab production starts being really constrained. A simple answer is while datacenters run continuously at close to max power, most chips currently produced are idle a lot of the time. Currently, smartphones are close to half of leading chip demand, but use a lot less energy per wafer area (trading transistors for serial operations and energy efficiency) and have low utilization as smartphones are mostly idle. The AI revolution means working our transistors way harder, dedicating them all to constantly-running, high-performance AI datacenters instead of idle, battery-powered/energy-saving devices. HT Carl Shulman for this point.

inference and multiple players) by 2030 will be a multiple of TSMC's current total leading-edge logic chip capacity, just for AI. TSMC ~doubled[66] in the past 5 years; they'd likely need to go ~at least twice as fast on their pace of expansion to meet AI chip demand. Massive new fab investments would be necessary.

Even if raw logic fabs won't be the constraint, chip-on-wafer-on-substrate (CoWoS) advanced packaging (connecting chips to memory, also made by TSMC, Intel, and others) and HBM memory (for which demand is enormous) are already key bottlenecks for the current AI GPU scaleup; these are more specialized to AI, unlike the pure logic chips, so there's less pre-existing capacity. In the near term, these will be the primary constraint on churning out more GPUs, and these will be the huge constraints as AI scales. Still, these are comparatively "easy" to scale; it's been incredible watching TSMC literally build "greenfield" fabs (i.e. entirely new facilities from scratch) to massively scale up CoWoS production this year (and Nvidia is even starting to find CoWoS alternatives to work around the shortage).

A new TSMC Gigafab (a technological marvel) costs around $20B in capex and produces 100k wafer-starts a month. For hundreds of millions of AI GPUs a year by the end of the decade, TSMC would need to build dozens of these—as well as a huge buildout for memory, advanced packaging, networking, etc., which will be a major fraction of capex. It could add up to over $1T of capex. It will be intense, but doable. (Perhaps the biggest roadblock will not be feasibility, but TSMC not even trying—TSMC does not yet seem AI-scaling-pilled! They think AI will "only" grow at a glacial 50% CAGR.)

Recent USG efforts like the CHIPS Act have been trying to onshore more AI chip production to the US (as insurance in case of the Taiwan contingency). While onshoring more of AI chip production to the US would be nice, it's less critical than having the actual datacenter (on which the AGI lives) in the US. If having chip production abroad is like having uranium deposits abroad, having the AGI datacenter abroad is like having the literal nukes be built and stored abroad. Given the dysfunction

and cost we've seen from building fabs in the US in practice, my guess is we should prioritize datacenters in the US while betting more heavily on democratic allies like Japan and South Korea for fab projects—fab buildouts there seem much more functional.

The Clusters of Democracy

BEFORE THE DECADE IS OUT, many trillions of dollars of compute clusters will have been built. The only question is whether they will be built in America. Some are rumored to be betting on building them elsewhere, especially in the Middle East. Do we really want the infrastructure for the Manhattan Project to be controlled by some capricious Middle Eastern dictatorship?

The clusters that are being planned today may well be the clusters AGI and superintelligence are trained and run on, not just the "cool-big-tech-product clusters." The national interest demands that these are built in America (or close democratic allies). Anything else creates an irreversible security risk: it risks the AGI weights getting stolen[67] (and perhaps be shipped to China); it risks these dictatorships physically seizing the datacenters (to build and run AGI themselves) when the AGI race gets hot; or even if these threats are only wielded implicitly, it puts AGI and superintelligence at unsavory dictator's whims. America sorely regretted her energy dependence on the Middle East in the 70s, and we worked so hard to get out from under their thumbs. We cannot make the same mistake again.

The clusters can be built in the US, and we have to get our act together to make sure it happens in the US. American national security must come first, before the allure of free-flowing Middle Eastern cash, arcane regulation, or even, yes, admirable climate commitments. We face a real system competition— can the requisite industrial mobilization only be done in "top-down" autocracies? If American business is unshackled, America can build like none other (at least in red states). Being willing to use natural gas, or at the very least a broad-based deregulatory agenda—NEPA exemptions, fixing FERC and

[67] It's a lot easier to do side-channel attacks to exfiltrate weights with physical access!

transmission permitting at the federal level, overriding utility regulation, using federal authorities to unlock land and rights of way—is a national security priority.

IN ANY CASE: the exponential is in full swing now.

In the "old days," when AGI was still a dirty word, some colleagues and I used to make theoretical economic models of what the path to AGI might look like. One feature of these models used to be a hypothetical "AI wakeup" moment, when the world started realizing how powerful these models could be and began rapidly ramping up their investments—culminating in multiple % of GDP towards the largest training runs.

It seemed far-off then, but that time has come. 2023 *was* "AI wakeup."[68] Behind the scenes, the most staggering techno-capital acceleration has been put into motion.

[68] I distinctly remember writing "THE TAKEOFF HAS STARTED" on my whiteboard in March of 2023.

Brace for the G-forces.

(What all of this means for NVDA/TSM/etc I leave as an exercise for the reader. Hint: Those with situational awareness bought much lower than you, but it's still not even close to fully priced in.)[69]

[69] Mainstream sell-side analysts seem to assume only 10-20% year-over-year growth in Nvidia revenue from CY24 to CY25, maybe $120B-$130B in CY25 (or at least did until very recently). Insane! It's been pretty obvious for a while that Nvidia is going to do over $200B of revenue in CY25.

IIIb. Lock Down the Labs: Security for AGI

The nation's leading AI labs treat security as an afterthought. Currently, they're basically handing the key secrets for AGI to the CCP on a silver platter. Securing the AGI secrets and weights against the state-actor threat will be an immense effort, and we're not on track.

> They met in the evening in Wigner's office. "Szilard outlined the Columbia data," Wheeler reports, "and the preliminary indications from it that at least two secondary neutrons emerge from each neutron-induced fission. Did this not mean that a nuclear explosive was certainly possible?" Not necessarily, Bohr countered.
>
> "We tried to convince him," Teller writes, "that we should go ahead with fission research but we should not publish the results. We should keep the results secret, lest the Nazis learn of them and produce nuclear explosions first."
>
> "Bohr insisted that we would never succeed in producing nuclear energy and he also insisted that secrecy must never be introduced into physics."
>
> *The Making of the Atomic Bomb*, p. 430

ON THE CURRENT COURSE, the leading Chinese AGI labs won't be in Beijing or Shanghai—they'll be in San Francisco and London. In a few years, it will be clear that the AGI secrets are the United States' most important national defense secrets—deserving treatment on par with B-21 bomber or Columbia-class submarine blueprints, let alone the proverbial

"nuclear secrets"—but today, we are treating them the way we would random SaaS software. At this rate, we're basically just handing superintelligence to the CCP.

All the trillions we will invest, the mobilization of American industrial might, the efforts of our brightest minds—none of that matters if China or others can simply steal the model *weights* (all a finished AI model is, all AGI will be, is a large file on a computer) or key *algorithmic secrets* (the key technical breakthroughs necessary to build AGI).

America's leading AI labs self-proclaim to be building AGI: they believe that the technology they are building will, before the decade is out, be the most powerful weapon America has ever built. But they do not treat it as such. They measure their security efforts against "random tech startups," not "key national defense projects." As the AGI race intensifies—as it becomes clear that superintelligence will be utterly decisive in international military competition—we will have to face the full force of foreign espionage. Currently, labs are barely able to defend against scriptkiddies, let alone have "North Korea-proof security," let alone be ready to face the Chinese Ministry of State Security bringing its full force to bear.

And this won't just matter years in the future. Sure, who cares if GPT-4 weights are stolen—what really matters in terms of weight security is that we can secure the AGI weights down the line, so we have a few years, you might say. (Though if we're building AGI in 2027, we really have to get moving!) But the AI labs are developing the *algorithmic secrets*—the key technical breakthroughs, the blueprints so to speak—for the AGI *right now* (in particular, the RL/self-play/synthetic data/etc "next paradigm" after LLMs to get past the data wall). AGI-level security for algorithmic secrets is necessary years before AGI-level security for weights. These algorithmic breakthroughs will matter more than a 10x or 100x larger cluster in a few years—this is a much bigger deal than export controls on compute, which the USG has been (presciently!) intensely pursuing. Right now, you needn't even mount a dramatic espionage operation to steal these secrets: just go to any SF party or look through the office windows.

Our failure today will be irreversible soon: in the next 12-24 months, we will leak key AGI breakthroughs to the CCP. It will be the national security establishment's single greatest regret before the decade is out.

The preservation of the free world against the authoritarian states is on the line—and a healthy lead will be the necessary buffer that gives us margin to get AI safety right, too. The United States has an advantage in the AGI race. But we will give up this lead if we don't get serious about security very soon. Getting on this, now, is maybe even the single most important thing we need to do today to ensure AGI goes well.

Underrate state actors at your peril

Too many smart people underrate espionage.

The capabilities of states and their intelligence agencies are extremely formidable. Even in normal, non-all-out-AGI-race times (and from the little we know publicly), nation-states (or less advanced actors) have been able to:

- Zero-click hack any desired iPhone and Mac with just a phone number,

- Infiltrate an airgapped atomic weapons program,

- Modify Google source code,

- Find dozens of zero-days a year that take on average of 7 years to detect,

- Spearfish major tech companies,

- Install keyloggers on employee devices,

- Insert trapdoors in encryption schemes,

- Steal information via electromagnetic emanations or vibration,

- Use just the noise from your computer to determine where you are on a video game map or steal a password,

- Gain direct access to sensitive systems like nuclear power plants,

- Exfiltrate 22 million security clearance files from the USG,

- Expose the financial information of 110 million customers by planting vulnerabilities in HVAC systems,

- Compromise computer hardware supply chains at large scale,

- Slip malicious code into updates to software dependencies used by top tech companies and the USG

- ... let alone planting spies or seducing, cajoling, or threatening employees (which happens effectively at large scales, but is less public)

- ... let alone special forces operations and similar (when things really get hot).

For a further taste of what we're dealing with when facing down intelligence agencies, I highly recommend *Inside the Aquarium*, a book by a Soviet GRU (military intelligence) defector.[70]

Already, China engages in widespread industrial espionage; the FBI director stated the PRC has a hacking operation greater than "every major nation combined." And just a couple months ago, the Attorney General announced the arrest of a Chinese national who had stolen key AI code from Google to take back with him to the PRC (back in 2022/23, and probably just the tip of the iceberg).[71]

But that's just the beginning. We must be prepared for our adversaries to "wake up to AGI" in the next few years. AI will become the #1 priority of every intelligence agency in the world. In that situation, they would be willing to employ extraordinary means and pay any cost to infiltrate the AI labs.

[70] One spoiler, as a taste: on graduation from their ~spy academy, before being sent overseas, aspiring spies had to prove their skills domestically: they had to acquire secret information from a Soviet scientist. The penalty for revealing state secrets, of course, was death. That is: graduating from the ~spy academy meant picking a countryman to condemn to death. HT Ilya Sutskever for the book recommendation.

[71] By the way, the indictment provides a great illustration of how easy it is to evade security at even Google, which likely has the best security of all the AI labs (given their ability to lean on Google's decades-long investment in security infrastructure). All it took to steal the code, without detection, was pasting code into Apple Notes, then exporting to pdf!

"DING exfiltrated these files by copying data from the Google source files into the Apple Notes application on his Google-issued MacBook laptop. DING then converted the Apple Notes into PDF files. and uploaded them from the Google network into DING Account 1. This method helped DING evade immediate detection." (From the indictment.)

He only got caught because he did a bunch of other stupid things, like immediately start prominent startups in China which got people suspicious (and later even came back to the US).

The threat model

THERE ARE TWO KEY ASSETS we must protect: model weights (especially as we get close to AGI, but which takes years of preparation and practice to get right) and algorithmic secrets (starting yesterday).

Model weights

An AI model is just a large file of numbers on a server. This can be stolen. All it takes an adversary to match your trillions of dollars and your smartest minds and your decades of work is to steal this file. (Imagine if the Nazis had gotten an exact duplicate of every atomic bomb made in Los Alamos.)

If we can't keep model weights secure, we're just building AGI for the CCP (and, given the current trajectory of AI lab security, even North Korea).

Even besides national competition, securing model weights is critical for preventing AI catastrophes as well. All of our hand-wringing and protective measures are for naught if a bad actor (say, a terrorist or rogue state) can just steal the model and do whatever they want with it, circumventing any safety layers. Whatever novel WMDs superintelligence could invent would rapidly proliferate to dozens of rogue states. Moreover, security is the first line of defense against uncontrolled or misaligned AI systems, too (how stupid would we feel if we failed to contain the rogue superintelligence because we didn't build and test it in an air-gapped cluster first?).

Securing model weights doesn't matter that much right now: stealing GPT-4 without the underlying recipe doesn't do that much for the CCP. But it will *really* matter in a few years, once we have AGI, systems that are genuinely incredibly powerful.

Perhaps the single scenario that most keeps me up at night is if China or another adversary is able to steal the automated-AI-researcher-model-weights on the cusp of an intelligence explosion. China could immediately use these to automate AI research themselves

(even if they had previously been way behind)—and launch their own intelligence explosion. That'd be all they need to automate AI research, and build superintelligence. Any lead the US had would vanish.

Moreover, this would immediately put us in an existential race; any margin for ensuring superintelligence is safe would disappear. The CCP may well try to race through an intelligence explosion as fast as possible—even months of lead on superintelligence could mean a decisive military advantage—in the process skipping all the safety precautions any responsible US AGI effort would hope to take. We would also have to race through the intelligence explosion to avoid complete CCP dominance. Even if the US still manages to barely pull out ahead in the end, the loss of margin would mean having to run enormous risks on AI safety.

We're miles away for sufficient security to protect weights today. Google DeepMind (perhaps the AI lab that has the best security of any of them, given Google infrastructure) at least straight-up admits this. Their Frontier Safety Framework outlines security levels 0, 1, 2, 3, and 4 (~1.5 being what you'd need to defend against well-resourced terrorist groups or cybercriminals, 3 being what you'd need to defend against the North Koreas of the world, and 4 being what you'd need to have even a shot of defending against priority efforts by the most capable state actors).[72] They admit to being at level 0 (only the most banal and basic measures). If we got AGI and superintelligence soon, we'd literally deliver it to terrorist groups and every crazy dictator out there!

Critically, developing the infrastructure for weight security probably takes many years of lead times—if we think AGI in ~3-4 years is a real possibility and we need state-proof weight security then, we need to be launching the crash effort *now*. Securing weights will require innovations in hardware and radically different cluster design; and security at this level can't be reached overnight, but requires cycles of iteration.

If we fail to prepare in time, our situation will be dire. We will be on the cusp of superintelligence, but years away from the se-

[72] Based off of their claimed correspondence of their security levels to RAND's weight security report's L1-L5

curity necessary. Our choice will be to press ahead, but directly deliver superintelligence to the CCP—with the existential race through the intelligence explosion that implies—or wait until the security crash program is complete, risking losing our lead.

Algorithmic secrets

While people are starting to appreciate (though not necessarily implement) the need for weight security, arguably even more important right now—and *vastly* underrated—is securing algorithmic secrets.

One way to think about this is that stealing the algorithmic secrets will be worth having a 10x or more larger cluster to the PRC:

- As discussed in Counting the OOMs, algorithmic progress is probably similarly as important as scaling up compute to AI progress. Given the baseline trend of ~0.5 OOMs of compute efficiency a year (+ additional algorithmic "unhobbling" gains on top), we should expect multiple OOMs-worth of algorithmic secrets between now and AGI. By default, I expect American labs to be years ahead; if they can defend their secrets, this could easily be worth 10x-100x compute.

  - (Note that we're willing to incur American investors 100s of billions of dollars of costs by export controlling Nvidia chips—perhaps a 3x increase in compute cost for Chinese labs—but we're leaking 3x algorithmic secrets all over the place!)

- Maybe even more importantly, *we may be developing the key paradigm breakthroughs for AGI right now.* As discussed previously, simply scaling up current models will hit a wall: the data wall. Even with way more compute, it won't be possible to make a better model. The frontier AI labs are furiously at work at what comes next, from RL to synthetic data. They will probably figure out some crazy stuff—essentially, the "AlphaGo self-play"-equivalent for general intelligence. Their inventions will be as key as the invention of the LLM paradigm originally was a number of years

ago, and they will be the key to building systems that go far beyond human-level. We still have an opportunity to deny China these key algorithmic breakthroughs, without which they'd be stuck at the data wall. But without better security in the next 12-24 months, we may well irreversibly supply China with these key AGI breakthroughs.

- It's easy to underrate how important an edge algorithmic secrets will be—because up until ~a couple years ago, everything was published. The basic idea was out there: scale up Transformers on internet text. Many algorithmic details and efficiencies were out there: Chinchilla scaling laws, MoE, etc. Thus, open source models today are pretty good, and a bunch of companies have pretty good models (mostly depending on how much $$$ they raised and how big their clusters are). But this will likely change fairly dramatically in the next couple years. Basically all of frontier algorithmic progress happens at labs these days (academia is surprisingly irrelevant), and the leading labs have stopped publishing their advances. We should expect far more divergence ahead: between labs, between countries, and between the proprietary frontier and open source models. A few American labs will be way ahead—a moat worth 10x, 100x, or more, way more than, say, 7nm vs. 3nm chips—unless they instantly leak the algorithmic secrets.[73]

Put simply, I think failing to protect algorithmic secrets is probably the most likely way in which China is able to stay competitive in the AGI race. (I discuss this more later.)

It's hard to overstate how bad algorithmic secrets security is right now. Between the labs, there are thousands of people with access to the most important secrets; there is basically no background-checking, silo'ing, controls, basic infosec, etc. Things are stored on easily hackable SaaS services. People gabber at parties in SF. Anyone, with all the secrets in their head, could be offered $100M and recruited to a Chinese lab at any point.[74] You can . . . just look through office windows. And so on. There are many articles, and rumors flying around SF, purporting to have extensive details of various lab algorithmic advances.

[73] I sometimes joke that AI lab algorithmic advances are not shared with the American research community, but they are being shared with the Chinese research community!

[74] Indeed, I've heard from friends that ByteDance emailed basically every person who was on the Google Gemini paper to recruit them, offering them L8 (a very senior position, with presumably similarly high pay), and pitching them by saying they'd report directly to the CTO in America of ByteDance.

AI lab security isn't much better than "random startup secu-
rity." Directly selling the AGI secrets to the CCP would at least
be more honest.

> ... Is this what we see at OpenAI or any other American AI
> lab? No. In fact, what we see is the opposite—the security
> equivalent of swiss cheese. Chinese penetration of these labs
> would be trivially easy using any number of industrial
> espionage methods, such as simply bribing the cleaning crew
> to stick USB dongles into laptops. My own assumption is that
> all such American AI labs are fully penetrated and that China
> is getting nightly downloads of all American AI research and
> code RIGHT NOW...
>
> *Marc Andreessen*

While it will be tough, I think these secrets are defensible.
There are probably only dozens of people who truly "need
to know" the key implementation details for a given algorith-
mic breakthrough at a given lab (even if a larger number need
to know the basic high-level idea)—you can vet, silo, and inten-
sively monitor these people, in addition to radically upgraded
infosec.

What "supersecurity" will require

There's a lot of low-hanging fruit on security at AI labs. Merely
adopting best practices from, say, secretive hedge funds or
Google-customer-data-level security, would put us in a much
better position with respect to "regular" economic espionage
from the CCP. Indeed, there are notable examples of private
sector firms doing remarkably well at preserving secrets. Take
quantitative trading firms (the Jane Streets of the world) for
example. A number of people have told me that in an hour of
conversation they could relay enough information to a com-
petitor such that their firm's alpha would go to ~zero—similar
to how many key AI algorithmic secrets could be relayed in a
short conversation—and yet these firms manage to keep these
secrets and retain their edge.

While most of America's leading AI labs have refused to put
the national interest first—rejecting even basic security mea-

sures in this tier, if they have any cost or require any prioritization of security—picking this low-hanging fruit would be well within their abilities.

But let's look out just a bit further. Once China begins to truly understand the import of AGI, we should expect the full force of their espionage efforts to come to bear; think billions of dollars invested, thousands of employees, and extreme measures (e.g., special operations strike teams) dedicated to infiltrating American AGI efforts. What will security for AGI and superintelligence require?

In short, *this will only be possible with government help*. Microsoft, for example, is regularly hacked by state actors (e.g., Russian hackers recently stole Microsoft executives' emails, as well as government emails Microsoft hosts). A high-level security expert working in the field estimated that even with a complete private crash course, China would still likely be able to exfiltrate the AGI weights if it was their #1 priority—the only way to get this probability to the single digits would require, more or less, a government project.

While the government does not have a perfect track record on security themselves, they're the only ones who have the infrastructure, know-how, and competencies to protect national-defense-level secrets. Basic stuff like the authority to subject employees to intense vetting; threaten imprisonment for leaking secrets; physical security for datacenters; and the vast know-how of places like the NSA and the people behind the security clearances (private companies simply don't have the expertise on state-actor attacks).

I'm not one of the people behind the security clearances, so I can't give a proper accounting of what security for AGI will truly require. The best public resource on this is RAND's report on weight security. To give a taste of what this state-actor proof security will actually mean:

- Fully airgapped datacenters, with physical security on par with most secure military bases (cleared personnel, physical fortifications, onsite response team, extensive surveillance

and extreme access control),

- – And not just for training clusters—inference clusters need the same intense security![75]

- Novel technical advances on confidential compute / hardware encryption[76] and extreme scrutiny on the entire hardware supply chain,

- All research personnel working from a SCIF (Sensitive Compartmented Information Facility, pronounced "skiff", see this visualization),

- Extreme personnel vetting and security clearances (including regular employee integrity testing and the like), constant monitoring and substantially reduced freedoms to leave, and rigid information siloing.

- Strong internal controls, e.g. multi-key signoff to run any code.

- Strict limitations on any external dependencies, and satisfying general requirements of TS/SCI networks.

- Ongoing intense pen-testing by the NSA or similar.

- And so on. . .

The giant AGI clusters are being mapped out, right now; the corresponding security effort must be too. If we are building AGI in just a few years, we have very little time.

Still, this immense effort shouldn't lead to fatalism. The saving grace on security is that the CCP probably isn't fully AGI-pilled yet, and so not yet investing in the most extreme efforts. American AI lab security "only" has to stay ahead of the curve compared to the intensity of Chinese espionage efforts. That means immediately upgrading security to stay ahead of "more normal" economic espionage (which we are far from resistant to, but private companies probably could be); and that means over the next couple years, as Chinese and other foreign espionage ramps up, rapidly upgrading to much more intense

[75] Inference fleets will likely be much larger than training clusters, and there will be overwhelming pressure to use these inference clusters to run automated AI researchers during the intelligence explosion (and run billions of superintelligences more broadly in the immediate aftermath). The AGI/superintelligence weights could thus be exfiltrated from these clusters as well. (I worry that this is underrated, and inference clusters will be much less protected.)

[76] But you can't rely only on this! Hardware encryption regularly gets side-channeled. Defense-in-depth is key, of course.

measures in cooperation with the government.

SOME ARGUE THAT strict security measures and their associated friction aren't worth it because they would slow down American AI labs too much. But I think that's mistaken:

- This is a tragedy of the commons problem. For a given lab's commercial interests, security measures that cause a 10% slowdown might be deleterious in competition with other labs. But the national interest is clearly better served if every lab were willing to accept the additional friction: American AI research is way ahead of Chinese and other foreign algorithmic progress, and America retaining 90%-speed algorithmic progress as our national edge is clearly better than retaining 0% as a national edge (with everything instantly stolen)!

- Moreover, ramping security now will be the *less* painful path in terms of research productivity in the long run. Eventually, inevitably, if only on the cusp of superintelligence, in the extraordinary arms race to come, the USG will realize the situation is unbearable and demand a security crackdown. It will be so much more painful, and cause much more of a slowdown, to have to implement extreme, state-actor-proof security measures from a standing start, rather than iteratively.

Others argue that even if our secrets or weights leak, we will still manage to eke out *just ahead* by being faster in other ways (so we needn't worry about security measures). That, too, is mistaken, or at least running *way* too much risk:

- As I discuss in a later piece, I think the CCP may well be able to brutely outbuild the US (a 100GW cluster will be much easier for them). More generally, China might not have the same caution slowing it down that the US will (both reasonable and unreasonable caution!). Even if stealing the algorithms or weights "only" puts them on par with the US model-wise, that might be enough for them to win the race to superintelligence.

- Moreover, even if the US squeaks out ahead in the end, the difference between a 1-2 year and 1-2 month lead will *really* matter for navigating the perils of superintelligence. A 1-2 year lead means at least a reasonable margin to get safety right, and to navigate the extremely volatile period around the intelligence explosion and post-superintelligence.[77] A mere 1-2 month lead means a breakneck international arms race with extreme pressures, racing through the intelligence explosion, and no room at all to get safety right. It is that neck-and-neck, existential race in which we face the greatest risks of self-destruction.

- Don't forget about Russia, Iran, North Korea, and so on. Their hacking capabilities are no slouch. On the current course, we're freely sharing superintelligence with them too! Without much better security, we're proliferating what will be our most powerful weapon to a plethora of incredibly dangerous, reckless, and unpredictable actors.[78]

[77] E.g., space to take an extra 6 months during the intelligence explosion for alignment research to make sure superintelligence doesn't go awry, time to stabilize the situation after the invention of some novel WMDs by directing these systems to focus on defensive applications, or simply time for human decision-makers to make the right decisions given an extraordinarily rapid pace of technological change with the advent of superintelligence.

[78] We try really hard to prevent nuclear proliferation to rogue states, even if we'd still be "ahead" on nuclear technology compared to their more limited arsenal, given the mayhem proliferation can cause.

We are not on track

WHEN IT FIRST became clear to a few that an atomic bomb was possible, secrecy, too, was perhaps the most contentious issue. In 1939 and 1940, Leo Szilard became known "throughout the American physics community as the leading apostle of secrecy in fission matters."[79] But he was rebuffed by most; secrecy was not at all something scientists were used to, and it ran counter to many of their basic instincts of open science. But it slowly became clear what had to be done: the military potential of this research was too great for it to simply be freely shared with the Nazis. And secrecy was finally imposed, just in time.

In the fall of 1940, Fermi had finished new carbon absorption measurements on graphite, suggesting graphite was a viable moderator for a bomb. Szilard assaulted Fermi with yet another secrecy appeal. "At this time Fermi really lost his temper; he really thought this was absurd," Szilard recounted. Luckily, further appeals were eventually successful, and Fermi reluctantly refrained from publishing his graphite results.

[79] *The Making of the Atomic Bomb, p. 509*

Figure 31: Leo Szilard.

At the same time, the German project had narrowed down on two possible moderator materials: graphite and heavy water. In early 1941 at Heidelberg, Walther Bothe made an incorrect measurement on the absorption cross-section of graphite, and concluded that graphite would absorb too many neutrons to sustain a chain reaction. Since Fermi had kept his result secret, the Germans did not have Fermi's measurements to check against, and to correct the error. *This was crucial: it left the German project to pursue heavy water instead—a decisive wrong path that ultimately doomed the German nuclear weapons effort.*[80]

[80] *The Making of the Atomic Bomb,* p. 507

If not for that last-minute secrecy appeal, the German bomb project may have been a much more formidable competitor— and history might have turned out very differently.

There's a real mental dissonance on security at the leading AI labs. They full-throatedly claim to be building AGI this decade. They emphasize that American leadership on AGI will be decisive for US national security. They are reportedly planning 7T chip buildouts that only make sense if you *really* believe in AGI. And indeed, when you bring up security, they nod and acknowledge "of course, we'll all be in a bunker" and smirk.

And yet the reality on security could not be more divorced from that. Whenever it comes time to make hard choices to prioritize security, startup attitudes and commercial interests prevail over the national interest. The national security advisor would have a mental breakdown if he understood the level of security at the nation's leading AI labs.

There are secrets being developed right now, that can be used for every training run in the future and will be the key unlocks to AGI, that are protected by the security of a startup and will be worth hundreds of billions of dollars to the CCP.[81] The reality is that, a) in the next 12-24 months, we will develop the key algorithmic breakthroughs for AGI, and promptly leak them to the CCP, and b) we are not even on track for our weights to be secure against rogue actors like North Korea, let alone an

[81] 100x+ compute efficiencies, when clusters worth $10s or $100s of billions are being built.

all-out effort by China, by the time we build AGI. "Good security for a startup" simply is not even close to good enough, and we have very little time before the egregious damage to the national security of the United States becomes irreversible.

We're developing the most powerful weapon mankind has ever created. The algorithmic secrets we are developing, right now, are literally the nation's most important national defense secrets—the secrets that will be at the foundation of the US and her allies' economic and military predominance by the end of the decade, the secrets that will determine whether we have the requisite lead to get AI safety right, the secrets that will determine the outcome of WWIII, the secrets that will determine the future of the free world. And yet AI lab security is probably worse than a random defense contractor making bolts.

It's madness.

Basically nothing else we do—on national competition, and on AI safety—will matter if we don't fix this, soon.

Figure 32: A billboard at the Oak Ridge, Tennessee uranium enrichment facilities, 1943.

IIIc. Superalignment

Reliably controlling AI systems much smarter than we are is an unsolved technical problem. And while it is a solvable problem, things could very easily go off the rails during a rapid intelligence explosion. Managing this will be extremely tense; failure could easily be catastrophic.

> The old sorcerer
> Has finally gone away!
> Now the spirits he controls
> Shall obey my commands.
> …
> I shall work wonders too.
> …
> Sir, I'm in desperate straits!
> *The spirits I summoned -*
> *I can't get rid of them.*
>
> Johann Wolfgang von Goethe
> "The Sorcerer's Apprentice"

By THIS POINT, you have probably heard of the AI doomers. You might have been intrigued by their arguments, or you might have dismissed them off-hand. You're loath to read yet another doom-and-gloom meditation.

I am not a doomer.[82] Misaligned superintelligence is probably not the biggest AI risk.[83] But I did spend the past year working on technical research on aligning AI systems as my day-job at OpenAI, working with Ilya and the Superalignment

[82] As not-very-politely certified by the doomer-in-chief! At the very least, I would call myself a strong optimist that this problem is solvable. I have spent considerable energies debating the AI pessimists and strongly advocated against policies like an AI pause.

[83] I'm most worried about *things just being totally crazy* around superintelligence, including things like novel WMDs, destructive wars, and unknown unknowns. Moreover, I think the arc of history counsels us to not underrate authoritarianism—and superintelligence might allow authoritarians to dominate for billions of years.

team. There is a very real technical problem: our current alignment techniques (methods to ensure we can reliably control, steer, and trust AI systems) won't scale to superhuman AI systems. What I want to do is explain what I see as the "default" plan for how we'll muddle through,[84] and why I'm optimistic. While not enough people are on the ball—we should have much more ambitious efforts to solve this problem!—overall, we've gotten lucky with how deep learning has shaken out, there's a lot of empirical low-hanging fruit that will get us part of the way, and we'll have the advantage of millions of automated AI researchers to get us the rest of the way.

But I also want to tell you why I'm worried. Most of all, ensuring alignment doesn't go awry will require extreme competence in managing the intelligence explosion. If we do rapidly transition from from AGI to superintelligence, we will face a situation where, *in less than a year*, we will go from recognizable human-level systems for which descendants of current alignment techniques will mostly work fine, to much more alien, vastly superhuman systems that pose a qualitatively different, fundamentally novel technical alignment problem; at the same time, going from systems where failure is low-stakes to extremely powerful systems where failure could be catastrophic; all while most of the world is probably going kind of crazy. It makes me pretty nervous.

By the time the decade is out, we'll have billions of vastly superhuman AI agents running around. These superhuman AI agents will be capable of extremely complex and creative behavior; we will have no hope of following along. We'll be like first graders trying to supervise with multiple doctorates.

In essence, we face a problem of handing off *trust*. By the end of the intelligence explosion, we won't have any hope of understanding what our billion superintelligences are doing (except as they might choose to explain to us, like they might to a child). And we don't yet have the technical ability to reliably guarantee even basic side constraints for these systems, like "don't lie" or "follow the law" or "don't try to exfiltrate your server". Reinforcement from human feedback (RLHF) works very well for adding such side constraints for current

[84] As Tyler Cowen says, muddling through is underrated!

systems—but RLHF relies on humans being able to understand and supervise AI behavior, which fundamentally won't scale to superhuman systems.

Simply put, without a very concerted effort, we won't be able to guarantee that superintelligence won't go rogue (and this is acknowledged by many leaders in the field). Yes, it may all be fine by default. But we simply don't know yet. Especially once future AI systems aren't just trained with imitation learning, but large-scale, long-horizon RL (reinforcement learning), they will acquire unpredictable behaviors of their own, shaped by a trial-and-error process (for example, they may learn to lie or seek power, simply because these are successful strategies in the real world!).

The stakes will be high enough that hoping for the best simply isn't a good enough answer on alignment.

The problem

The superalignment problem

We've been able to develop a very successful method for aligning (i.e., steering/controlling) *current* AI systems (AI systems dumber than us!): Reinforcement Learning from Human Feedback (RLHF). The idea behind RLHF is simple: the AI system tries stuff, humans rate whether its behavior was good or bad, and then reinforce good behaviors and penalize bad behaviors. That way, it learns to follow human preferences.

Indeed, RLHF has been the key behind the success of ChatGPT and others.[85] Base models had lots of raw smarts but weren't applying these in a useful way by default; they usually just responded with a garbled mess resembling random internet text. Via RLHF, we can steer their behavior, instilling important basics like instruction-following and helpfulness. RLHF also allows us to bake in safety guardrails: for example, if a user asks me for bioweapon instructions, the model should probably refuse.[86]

[85] Ironically, the safety guys made the biggest breakthrough for enabling AI's commercial success by inventing RLHF! Base models had lots of raw smarts but were unsteerable and thus unusable for most applications.

[86] This highlights an important distinction: the *technical ability* to align (steer/control) a model is separate from a values question of *what to align to*. There have been many political controversies about the latter question. And while I agree with the opposition to some of the outgrowths here, that shouldn't distract from the basic technical problem. Yes, alignment techniques can be misused—but we will need better alignment techniques to ensure even basic side constraints for future models, like follow instructions or follow the law. See also "AI alignment is distinct from its near-term applications".

The core technical problem of superalignment is simple: how do we control AI systems (much) smarter than us?

RLHF will predictably break down as AI systems get smarter, and we will face fundamentally new and qualitatively different technical challenges. Imagine, for example, a superhuman AI system generating a million lines of code in a new programming language it invented. If you asked a human rater in an RLHF procedure, "does this code contain any security backdoors?" they simply wouldn't know. They wouldn't be able to rate the output as good or bad, safe or unsafe, and so we wouldn't be able to reinforce good behaviors and penalize bad behaviors with RLHF.

Figure 33: Aligning AI systems via human supervision (as in RLHF) won't scale to superintelligence.
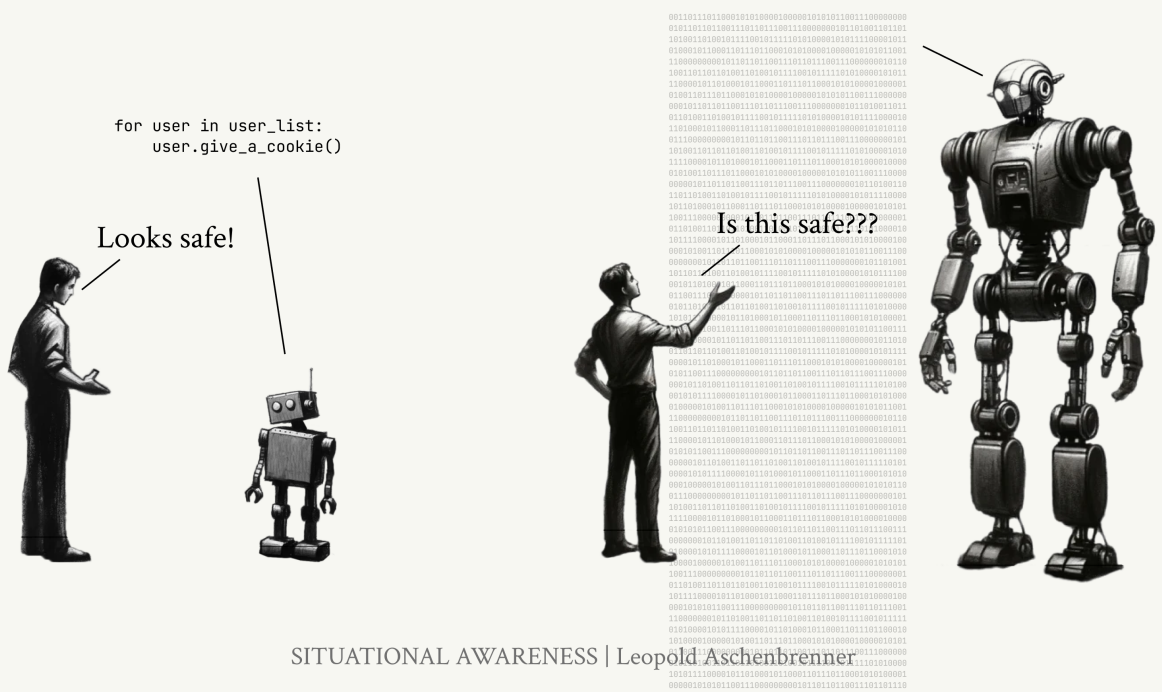
Even now, AI labs already need to pay expert software engineers to give RLHF ratings for ChatGPT code—the code current models can generate is already pretty advanced! Human labeler-pay has gone from a few dollars for MTurk labelers to

~\$100/hour for GPQA questions[87] in the last few years. In the (near) future, even the best human experts spending lots of time won't be good enough. We're starting to hit early versions of the superalignment problem in the real world now, and very soon this will be a major issue even just for practically deploying next-generation systems. It's clear we will need a successor to RLHF that scales to AI capabilities better than human-level, where human supervision breaks down. In some sense, the goal of superalignment research efforts is to repeat the success story of RLHF: make the basic research bets that will be necessary to steer and deploy AI systems a couple years down the line.

[87] "We estimate an average hourly payment of approximately \$95 per hour," p.3 of the GPQA paper.

What failure looks like

People too often just picture a "GPT-6 chatbot," informing their intuitions that surely these wouldn't be dangerously misaligned. As discussed previously in this series, the "unhobbling" trajectory points to agents, trained with RL, in the near future. I think Roger's graphic gets it right.
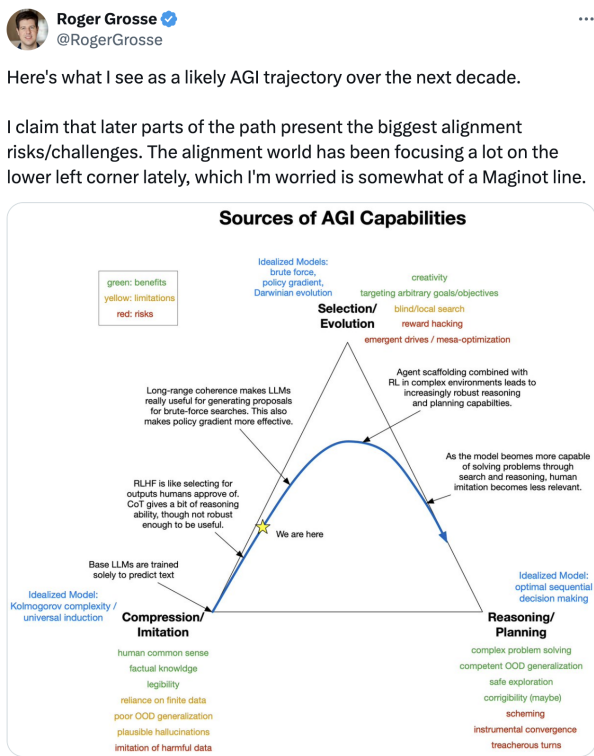
Figure 34: Roger Grosse on the evolution of the alignment problem as systems become more advanced.

One way to think of what we're trying to accomplish with alignment, from the safety perspective, is add *side-constraints.* Consider a future powerful "base model" that, in a second stage of training, we train with long-horizon RL to run a business and make money[88] (as a simplified example):

- By default, it may well learn to lie, to commit fraud, to deceive, to hack, to seek power, and so on—simply because these can be successful strategies to make money in the real world![89]

- What we want is to add side-constraints: don't lie, don't break the law, etc.

- But here we come back to the fundamental issue of aligning superhuman systems: we won't be able to understand what they are doing, and so we won't be able to notice and penalize bad behavior with RLHF.[90]

If we can't add these side-constraints, it's not clear what will happen. *Maybe* we'll get lucky and things will be benign by default (for example, maybe we can get pretty far without the AI systems having long-horizon goals, or the undesirable behaviors will be minor). But it's also totally plausible they'll learn much more serious undesirable behaviors: they'll learn to lie, they'll learn to seek power, they'll learn to behave nicely when humans are looking and pursue more nefarious strategies when we aren't watching, and so on.

The superalignment problem being unsolved means that we simply won't have the ability to ensure even these basic side constraints for these superintelligence systems, like *"will they reliably follow my instructions?"* or *"will they honestly answer my questions?"* or *"will they not deceive humans?"*. People often associate alignment with some complicated questions about human values, or jump to political controversies, but deciding on what behaviors and values to instill in the model, while important, is a separate problem. The primary problem is that for whatever you want to instill the model (including ensuring very basic things, like "follow the law"!) we don't yet know how to do that for the very powerful AI systems we are building very

[88] Very simplified, think of an AI system trying, via trial and error, to maximize money over a period of a year, and the final, trained AI model being the result of a selection process that selects for the AI systems that were most successful at maximizing money.

[89] What RL is doing is simply exploring strategies for succeeding at the objective. If a strategy works, it is reinforced in the model. So if lying, fraud, power-seeking, etc. (or patterns of thinking that could lead to these sorts of behaviors in at least some situations) work, these will also be reinforced in the model.

[90] (or inference-time monitoring models trained with human supervision)

soon.

Again, the consequences of this aren't totally clear. What is clear is that superintelligence will have vast capabilities—and so misbehavior could fairly easily be catastrophic. What's more, I expect that within a small number of years, these AI systems will be integrated in many critical systems, including military systems (failure to do so would mean complete dominance by adversaries). It sounds crazy, but remember when everyone was saying we wouldn't connect AI to the internet? The same will go for things like "we'll make sure a human is always in the loop!"—as people say today.

Alignment failures then might look like isolated incidents, say, an autonomous agent committing fraud, a model instance self-exfiltrating, an automated researcher falsifying an experimental result, or a drone swarm overstepping rules of engagement. But failures could also be much larger scale or more systematic—in the extreme, failures could look more like a robot rebellion. We'll have summoned a fairly alien intelligence, one *much* smarter than us, one whose architecture and training process wasn't even designed by us but some supersmart previous generation of AI systems, one where we can't even begin to understand what they're doing, it'll be running our military, and its goals will have been learned by a natural-selection-esque process.

Unless we solve alignment—unless we figure out how to instill those side-constraints—there's no particular reason to expect this small civilization of superintelligences will continue obeying human commands in the long run. It seems totally within the realm of possibilities that at some point they'll simply conspire to cut out the humans, whether suddenly or gradually.

The intelligence explosion makes this all incredibly tense
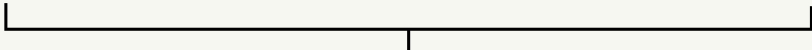
I am optimistic that superalignment is a solvable technical problem. Just like we developed RLHF, so we can develop the successor to RLHF for superhuman systems and do the science that gives us high confidence in our methods. If things continue to progress iteratively, if we insist on rigorous safety

testing and so on, it should all be doable (and I'll discuss my current best-guess of how we'll muddle through more in a bit).

What makes this incredibly hair-raising is the possibility of an intelligence explosion: that we might make the transition from roughly human-level systems to *vastly* superhuman systems extremely rapidly, perhaps in less than a year.

Alignment during the intelligence explosion

	AGI	**Superintelligence**
Required alignment technique	RLHF++	Novel, qualitatively different technical solutions
Failures	Low-stakes	Catastrophic
Architecture and algorithms	Familiar, descendants of current systems, fairly benign safety properties	Alien. Designed by previous-generation super-smart AI system
Backdrop	World is normal	World is going crazy, extraordinary pressures
Epistemic state	We can understand what the systems are doing, how they work, and whether they're aligned.	We have no ability to understand what's going on, how to tell if systems are still aligned and benign, what the systems are doing, and we are entirely reliant on trusting the AI systems.

Transition in <1 year?
with very little time to get decisions right

SITUATIONAL AWARENESS | Leopold Aschenbrenner

Figure 35: The intelligence explosion makes superalignment incredibly hair-raising.

- We will extremely rapidly go from systems where RLHF works fine—to systems where it will totally break down. This leaves us extremely little time to iteratively discover and

address ways in which our current methods will fail.

- At the same time, we will extremely rapidly go from systems where failures are fairly low-stakes (ChatGPT said a bad word, so what)—to extremely high-stakes (oops, the superintelligence self-exfiltrated from our cluster, now it's hacking the military). Rather than iteratively encountering increasingly more dangerous safety failures in the wild, the first notable safety failures we encounter might already be catastrophic.

- The superintelligence we get by the end of it will be *vastly* superhuman. We'll be entirely reliant on trusting these systems, and trusting what they're telling us is going on—since we'll have no ability of our own to pierce through what exactly they're doing anymore.

- The superintelligence we get by the end of it could be quite *alien*. We'll have gone through a decade or more of ML advances during the intelligence explosion, meaning the architectures and training algorithms will be totally different (with potentially much riskier safety properties).

  - One example that's very salient to me: we may well bootstrap our way to human-level or somewhat-superhuman AGI with systems that reason via chains of thoughts, i.e. via English tokens. This is *extraordinarily* helpful, because it means the models "think out loud" letting us catch malign behavior (e.g., if it's scheming against us). But *surely* having AI systems think in tokens is not the most efficient means to do it, *surely* there's something much better that does all of this thinking via internal states—and so the model by the end of the intelligence explosion will almost certainly *not* think out loud, i.e. will have completely uninterpretable reasoning.

- This will be an incredibly volatile period, potentially with the backdrop of an international arms race, tons of pressure to go faster, wild new capabilities advances every week with basically no human-time to make good decisions, and so on. We'll face tons of ambiguous data and high-stakes decisions.

– Think: "We caught the AI system doing some naughty things in a test, but we adjusted our procedure a little bit to hammer that out. Our automated AI researchers tell us the alignment metrics look good, but we don't really understand what's going on and don't fully trust them, and we don't have any strong scientific understanding that makes us confident this will continue to hold for another couple OOMs. So, we'll *probably* be fine? Also China just stole our weights and they're launching their own intelligence explosion, they're right on our heels."

It just really seems like this could go off the rails. To be honest, it sounds terrifying.

Yes, we will have AI systems to help us. Just like they'll automate capabilities research, we can use them to automate alignment research. That will be key, as I discuss below. But—can you trust the AI systems? You weren't sure whether they were aligned in the first place—are they actually being honest with you about their claims about alignment science? Will automated alignment research be able to keep up with automated capabilities research (for example, because automating alignment is harder, e.g. because there are less clear metrics we can trust compared to improving model capabilities, or there's a lot of pressure to go full-speed on capabilities progress because of the international race)? And AI won't be able to fully substitute for the still-human decision makers making good calls in this incredibly high-stakes situation.

The default plan: how we can muddle through

I THINK we can harvest wins across a number of empirical bets, which I'll describe below, to align somewhat-superhuman systems. Then, if we're confident we can trust these systems, we'll need to use these somewhat-superhuman systems to automate alignment research—alongside the automation of AI research in general, during the intelligence explosion—to figure out how to solve alignment to go the rest of the way.

Aligning somewhat-superhuman models

Aligning human-level systems won't be enough. Even the first systems that can do automated AI research, i.e. start the intelligence explosion, will likely already be substantially superhuman in many domains. This is because AI capabilities are likely to be somewhat spikey—by the time AGI is human-level at whatever a human AI researcher/engineer is worst at, it'll be superhuman at many other things. For example, perhaps the ability for AI systems to effectively coordinate and plan lags behind, meaning that by the time the intelligence explosion is in full force they'll probably already be superhuman coders, submitting million-line pull requests in new programming languages they devised, and they'll be superhuman at math and ML.

These early-intelligence-explosion-systems will start being quantitatively and qualitatively superhuman, at least in many domains. But they'll look much closer to the systems we have today in terms of architecture, and the intelligence gap we need to cover is much more manageable. (Perhaps if humans trying to align true superintelligence is like a first grader trying to supervise a PhD graduate, this is more like a smart high schooler trying to supervise a PhD graduate.)

More generally, the more we can develop good science now, the more we'll be in a position to verify that things aren't going off the rails during the intelligence explosion. Even having good metrics we can trust for superalignment is surprisingly difficult—but without reliable metrics during the intelligence explosion, we won't know whether pressing on is safe or not.

Here are some of the main research bets I see for crossing the gap between human-level and somewhat-superhuman systems.

EVALUATION IS EASIER THAN GENERATION.   We get some of the way "for free," because it's easier for us to evaluate outputs (especially for egregious misbehaviors) than it is to generate them ourselves. For example, it takes me months or years of hard work to write a paper, but only a couple hours to tell if a paper someone has written is any good (though perhaps longer

to catch fraud). We'll have teams of expert humans spend a lot of time evaluating every RLHF example, and they'll be able to "thumbs down" a lot of misbehavior even if the AI system is somewhat smarter than them. That said, this will only take us so far (GPT-2 or even GPT-3 couldn't detect nefarious GPT-4 reliably, even though evaluation is easier than generation!)

SCALABLE OVERSIGHT. We can use AI assistants to help humans supervise other AI systems—the human-AI team being able to extend supervision farther than the human could alone. For example, imagine an AI system generates 1M lines of complex code. It's easier for a human to double-check that a given line of code has a bug if an AI assistant has pointed out that (say) line 394,894 looks fishy, than it is for a human to find that same bug from scratch. A model trained to critique the code written by another model could thus help humans supervise a system with narrowly superhuman coding abilities.

Several scalable oversight strategies have been proposed, including debate, market-making, recursive reward modeling, and prover-verifier games, as well as simplified versions of those ideas lik ecritiques. Models are now strong enough that it's possible to empirically test these ideas, making direct progress on scalable oversight.

I expect this to help a lot for the "quantitatively" superhuman part of the problem, such as the aforementioned million-lines of code example. But I am less optimistic that scalable oversight techniques can really help for the "qualitatively" superhuman part of the problem—imagine the model invents quantum physics when you only understand Newtonian physics.

GENERALIZATION. Even with scalable oversight, we won't be able to supervise AI systems on really hard problems, problems beyond human comprehension. However, we can study: how will the AI systems *generalize* from human supervision on easy problems (that we do understand and can supervise) to behave on the hard problems (that we can't understand and can no longer supervise)? For example, perhaps supervising a model to be honest in simple cases generalizes benignly to

the model just being honest in general, even in cases where it's doing extremely complicated things we don't understand.

There's a lot of reasons to be optimistic here: part of the magic of deep learning is that it often generalizes in benign ways (for example, RLHF'ing with only labels on English examples also tends to produce good behavior when it's speaking French or Spanish, even if that wasn't part of the training). I'm fairly optimistic that there will both be pretty simple methods that help nudge the models' generalization in our favor, and that we can develop a strong scientific understanding that helps us predict when generalization will work and when it will fail. To a greater extent that for scalable oversight, the hope is that this will help with alignment even in the "qualitatively" superhuman case.

Here's another way of thinking about this: if a superhuman model is misbehaving, say breaking the law, intuitively *the model should already know* that it's breaking the law. Moreover, "is this breaking the law" is probably a pretty natural concept to the model—and it will be salient in the model's representation space. The question then is: can we "summon" this concept from the model with only weak supervision?

I'm particularly partial to this direction (and perhaps biased), because I helped introduce this with some recent work with some colleagues at OpenAI. In particular, we studied an analogy to the problem of humans supervising superhuman systems— can a *small model* align a *larger (smarter) model*? We found that generalization does actually get you cross some (but certainly not all) of the intelligence gap between supervisor and supervisee, and that in simple settings there's a lot you can do to improve it.

INTERPRETABILITY.   One intuitively-attractive way we'd hope to verify and trust that our AI systems are aligned is if we could understand what they're thinking! For example, if we're worried that AI systems are deceiving us or conspiring against us, access to their internal reasoning should help us detect that.

By default, modern AI systems are inscrutable black boxes.
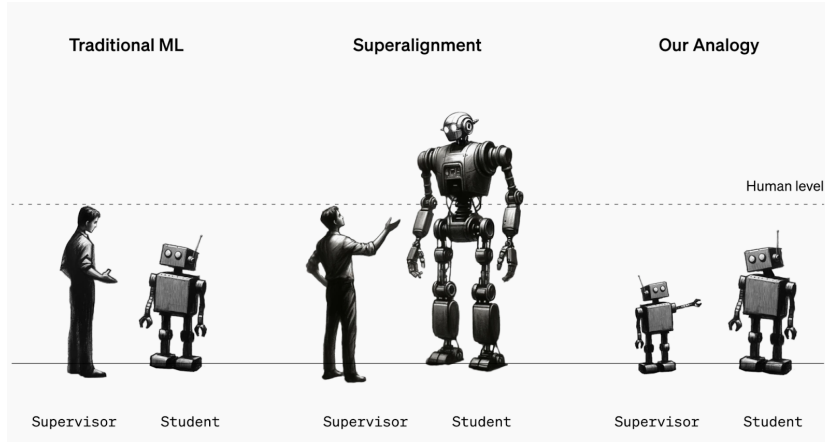
Figure 36: A simple analogy for studying superalignment: instead of a human supervising a superhuman model, we can study a small model supervising a large model. For example, can we align GPT-4 with only GPT-2 supervision? Will that result in GPT-4 appropriately generalizing "what GPT-2 meant"? From Weak-to-strong generalization.

Yet it seems like we should be able to do amazing "digital neuroscience"—after all, we have perfect access to model internals.

There's a few different approaches here, going from "most ambitious and 'cool' but will be very hard" to "hackier things that are easier and might just work".

Mechanistic interpretability.   Try to fully reverse-engineer large neural networks from the ground up—fully disentangle the inscrutable matrices, so to speak.

Chris Olah's team at Anthropic has done much of the pioneering work on this, starting by understanding simple mechanisms in very small models. There's incredibly exciting progress happening recently, and I'm thrilled about the overall level of activity in this space.

That said, I'm worried fully reverse-engineering superhuman AI systems will just be an intractable problem—similar, to, say "fully reverse engineering the human brain"—and I'd put this work mostly in the "ambitious moonshot for AI safety" rather than "default plan for muddling through" bucket.[91]

(Neel Nanda's 200 open problems in mechanistic interpretability also gives a flavor of this kind of research.)

[91] Moreover, even recent progress in mechanistic interpretability to "disentangle the features" of models with sparse autoencoders doesn't on its own solve the problem of how to deal with superhuman models. For one, the model might simply "think" in superhuman concepts you don't understand. Moreover, how do you know which feature is the one you want? You still don't have ground truth labels. For example, there might be tons of different features that look like the "truth feature" to you, one of which is "what the model actually knows" and the others being "what would xyz human think" or "what do the human raters want me to think" etc.

Sparse autoencoders won't be enough on their own, but they will be a tool—an incredibly helpful tool!—that ultimately has to cash out in helping with things like the science of generalization.

"Top-down" interpretability.   If mechanistic interpretability tries to reverse engineer neural networks "from the bottom up," other work takes a more targeted, "top-down" approach, trying to locate information in a model without full understanding of how it is processed.

For example, we might try to build an "AI lie detector" by identifying the parts of the neural net that "light up" when an AI system is lying. This can be a lot more tractable (even if it gives less strong guarantees).

Over the past couple of years, there's been a flurry of exciting work in this area. CCS is able to identify a "truth direction" in models with only unsupervised data.[92] ROME is able to identify the part of the model that knows the Eiffel Tower is in Paris—and then directly edit the model's knowledge to place the Eiffel Tower in Rome. Representation Engineering and Inference-time Interventions demonstrate using top-down techniques to detect lying and hallucinations and surgically control model behavior on jailbreaking, power-seeking, fairness, truthfulness, and more. There's other creative work on lie detection that doesn't even require model internals.

I'm increasingly bullish that top-down interpretability techniques will be a powerful tool—i.e., we'll be able to build something like an "AI lie detector"[93]—and without requiring fundamental breakthroughs in understanding neural nets.

Chain-of-thought interpretability.   As mentioned earlier, I think it's quite plausible that we'll bootstrap our way to AGI with systems that "think out loud" via chains of thought. Even if it's not the most efficient algorithm in the limit (I'd be very surprised if superintelligence still used English-chain-of-thought, rather than thinking via some sort of recurrent internal states), we may well still benefit from this property for the first AGIs. This would be an incredible boon for interpretability: we'd have access to the AGI's "internal monologue"! That would make it relatively easy to detect egregious alignment failures.

There's a ton of work to do here, however, if we wanted to rely on this. How do we ensure that the CoT remains *legible*? (It

[92] Essentially, needing only the consistency properties of truth, rather than strong/ground truth labels of true/false, which we won't have for superhuman systems.

[93] I'm still worried about the scalability of a lot of these techniques to very superhuman models—I think explicitly or implicitly they mostly rely on ground-truth labels, i.e. a supervisor smarter than the model, and/or favorable generalization.

may simply drift from understandable English to unintelligible model-speak, depending on how we e.g. use RL to train models—can we add some simple constraints to ensure it remains legible?) How do we ensure the CoT is *faithful*, i.e. actually reflects what models are thinking? (E.g, there's some work that shows in certain situations, models will just make up posthoc reasoning in their CoT that don't actually reflect their actual internal reasoning for an answer.)

My best guess is that some simple measurement of legibility and faithfulness, and some simple hacks to preserve legibility and faithfulness longer, could go quite far. Yes, this won't work in some worlds, and it's a bit of a simple hack, but it's such low-hanging fruit; this direction is criminally underrated in my view.

ADVERSARIAL TESTING AND MEASUREMENTS. Along the way, it's going to be critical to stress test the alignment of our systems at every step—our goal should be to encounter every failure mode in the lab before we encounter it in the wild. This will require substantially advancing techniques for automated red-teaming. For example, if we deliberately plant backdoors or misalignments into models, would our safety training would have caught and gotten rid of them? (Early work suggests that "sleeper agents" can survive through safety training, for example.)

More generally, it'll be critical to have good measurements of alignment. Does the model have the power to be misaligned? For example, does it have long-horizon goals, and what sorts of drives it is learning? And what are clear "red lines"? For example, a very intuitive bound might be "model reasoning (chain of thoughts) always has to remain legible and faithful." (As Eric Schmidt says, the point at which AI agents can talk to each other in a language we can't understand, we should unplug the computers.) Another might be developing better measurements for whether models are being fully honest.

The science of measuring alignment is still in its infancy; improving this will be critical for helping us make the right trade-offs on risk during the intelligence explosion. Doing the science

that lets us measure alignment and gives us an understanding of "what evidence would be sufficient to assure us that the next OOM into superhuman territory is safe?" is among the very-highest priority work for alignment research today (beyond just work that tries to extend RLHF further to "somewhat-superhuman" systems).

Automating alignment research

Ultimately, we're going to need to automate alignment research. There's no way we'll manage to solve alignment for true superintelligence directly; covering that vast of an intelligence gap seems extremely challenging. Moreover, by the end of the intelligence explosion—after 100 million automated AI researchers have furiously powered through a decade of ML progress—I expect much more alien systems in terms of architecture and algorithms compared to current system (with potentially less benign properties, e.g. on legibility of CoT, generalization properties, or the severity of misalignment induced by training).

But we also don't have to solve this problem just on our own. If we manage to align somewhat-superhuman systems enough to trust them, we'll be in an incredible position: we'll have millions of automated AI researchers, smarter than the best AI researchers, at our disposal. Leveraging these army of automated researchers properly to solve alignment for even-more superhuman systems will be decisive.

(This applies more generally, by the way, for the full spectrum of AI risks, including misuse and so on. The best route—perhaps the only route—to AI safety in all of these cases, will involve properly leveraging early AGIs for safety; for example, we should put a bunch of them to work on automated research to improve security against foreign actors exfiltrating weights, others on shoring up defenses against worst-case bioattacks, and so on.)

Getting automated alignment right during the intelligence explosion will be extraordinarily high-stakes: we'll be going through many years of AI advances in mere months, with little

human-time to make the right decisions, and we'll start entering territory where alignment failures could be catastrophic. Labs should be willing to commit a large fraction of their compute to automated alignment research (vs. automated capabilities research) during the intelligence explosion, if necessary. We'll need strong guarantees that let us trust the automated alignment research being produced and much better measurements for misalignment than we have today to know whether we're still in the clear. For every OOM we want to ascend, we will need to have extremely high confidence in our alignment approaches for that next OOM.

Most of all, getting this right will require an extreme degree of competence, seriousness, and willingness to make hard trade-offs. Once we're approaching true superintelligence, the stakes get very real, and failure—rogue AI—could be catastrophic. There will be moments with ambiguous evidence, when metrics will superficially look ok, but there are some warning signs that hint at looming danger, and we mostly don't have a clue what's going on anymore. Navigating this situation responsibly will require making very costly tradeoffs. Eventually, the decision to greenlight the next generation of superintelligence will need to be taken as seriously as the decision to launch a military operation.

Superdefense

"Getting alignment right" should only be the first of many layers of defense during the intelligence explosion. Alignment will be hard; there *will* be failures along the way. If at all possible, we need to be in a position where alignment can fail—but failure wouldn't be catastrophic. This could mean:

- *Security.* An airgapped cluster is the first layer of defense against superintelligence attempting to self-exfiltrate and doing damage in the real world. And that's only the beginning; we'll need much more extreme security against model self-exfiltration[94] across the board, from hardware encryption to many-key signoff.[95] And once our AGIs can do automated

[94] A model stealing its own weights, to make copies of itself outside of the original datacenter.

[95] protecting against humans fooled or persuaded by the AI to help it exfiltrate

R&D, we should put them to work on further hardening security as well.

- *Monitoring.* If our AI systems are up to something fishy or malevolent—or a rogue employee tries to use them for unauthorized activities—we need to be able to catch it. However, this is not a trivial task: we will have millions of AGI instances running our cluster, executing complex code and so on. We will need advanced monitoring systems that leverage the models themselves, including monitoring model outputs, chains of thought, model-written code, server/network activity, and interactions with researchers/other humans; essentially, leveraging some AI systems to nark on other AI systems. Moreover, we'll want to develop and employ sophisticated "control" protocols, using less powerful but trusted models to protect against subversion from more powerful but less trusted models. This may require, for example, dedicating the same amount of compute to monitoring model instances as to compute for active model instances.

- *Targeted capability limitations.* As much as possible, we should try to limit the model's capabilities in targeted ways that reduce fallout from failure. A central example of this might be scrubbing everything related to biology and chemistry from model training (or using "unlearning" techniques); this helps cut off paths to catastrophic real world damage that occur via biological and chemical weapons

- *Targeted training method restrictions.* There are likely some ways of training models that are inherently riskier—more likely to produce severe misalignments—than others. For example, imitation learning seems relatively safe (hard to see how that would lead to models that have dangerous long-term internal goals), while we should avoid long-horizon outcome-based RL (which seems much more likely to lead to the model learning undesirable long-term goals) as long as possible. We should also avoid training against our interpretability methods and our monitoring setup (such that our models aren't simply trained to work around these). There might be additional desirable limitations, for example, maintaining legible and faithful chains of thought as long as

possible. We should define these constraints ahead of time, maintain them as long as we can throughout the intelligence explosion, and only drop them if absolutely necessary.

• And there's likely a lot more possible here.

Will these be foolproof? Not at all. True superintelligence is likely able to get around most-any security scheme for example. Still, they buy us a lot more margin for error—and we're going to need any margin we can get. We'll want to use that margin to get in a position where we have very high confidence in our alignment techniques, only relaxing "superdefense" measures (for example, deploying the superintelligence in non-airgapped environments) concomitant with our confidence.

Things will get dicey again once we move to deploying these AI systems in less-controlled settings, for example in military applications. It's likely circumstances will force us to do this fairly quickly, but we should always try to buy as much margin for error as much as possible—for example, rather than just directly deploying the superintelligences "in the field" for military purposes, using them to do R&D in a more isolated environment, and only deploying the specific technologies they invent (e.g., more limited autonomous weapons systems that we're more confident we can trust).

Why I'm optimistic, and why I'm scared

I'm INCREDIBLY BULLISH on the technical tractability of the superalignment problem. It feels like there's tons of low-hanging fruit everywhere in the field. More broadly, the empirical realities of deep learning have shaken out more in our favor compared to what some speculated 10 years ago. For example, deep learning *generalizes* surprisingly benignly in many situations: it often just "does the thing we meant" rather than picking up some abstruse malign behavior.[96] Moreover, while fully understanding model internals will be hard, at least for the initial AGIs we have a decent shot of interpretability—we can get them to reason transparently via chains of thought, and

[96] Though, of course, while that may apply to current models and perhaps human-level systems, we have to be careful about attempting to extrapolate evidence from current models to future vastly superhuman models.

hacky techniques like representation engineering work surprisingly well as "lie detectors" or similar.

I think there's a pretty reasonable shot that "the default plan" to align "somewhat-superhuman" systems will mostly work.[97] Of course, it's one thing to speak about a "default plan" in the abstract—it's another if the team responsible for executing that plan is you and your 20 colleagues (much more stressful!).[98] There's still an incredibly tiny number of people seriously working on solving this problem, maybe a few dozen serious researchers. Nobody's on the ball! There's so much interesting and productive ML research to do on this, and the gravity of the challenge demands a much more concerted effort than what we're currently putting forth.

But that's just the first part of the plan—what really keeps me up at night is the intelligence explosion. Aligning the first AGIs, the first somewhat-superhuman systems, is one thing. Vastly superhuman, alien superintelligence is a new ballgame, and it is a *scary* ballgame.

The intelligence explosion will be more like running a war than launching a product. We're not on track for superdefense, for an airgapped cluster or any of that; I'm not sure we would even realize if a model self-exfiltrated. We're not on track for a sane chain of command to make any of these insanely high-stakes decisions, to insist on the very-high-confidence appropriate for superintelligence, to make the hard decisions to take extra time before launching the next training run to get safety right or dedicate a large majority of compute to alignment research, to recognize danger ahead and avert it rather than crashing right into it. Right now, no lab has demonstrated much of a willingness to make any costly tradeoffs to get safety right (we get lots of safety committees, yes, but those are pretty meaningless). By default, we'll probably stumble into the intelligence explosion and have gone through a few OOMs before people even realize what we've gotten into.

We're counting way too much on luck here.

[97] To be clear, given the stakes, I think "muddling through" is in some sense a terrible plan. But it might be all we've got.

[98] I'm reminded of Scott Aaronson's letter to his younger self:

"There's a company building an AI that fills giant rooms, eats a town's worth of electricity, and has recently gained an astounding ability to converse like people. It can write essays or poetry on any topic. It can ace college-level exams. It's daily gaining new capabilities that the engineers who tend to the AI can't even talk about in public yet. Those engineers do, however, sit in the company cafeteria and debate the meaning of what they're creating. What will it learn to do next week? Which jobs might it render obsolete? Should they slow down or stop, so as not to tickle the tail of the dragon? But wouldn't that mean someone else, probably someone with less scruples, would wake the dragon first? Is there an ethical obligation to tell the world more about this? Is there an obligation to tell it less?

I am—you are—spending a year working at that company. My job—your job—is to develop a mathematical theory of how to prevent the AI and its successors from wreaking havoc. Where "wreaking havoc" could mean anything from turbocharging propaganda and academic cheating, to dispensing bioterrorism advice, to, yes, destroying the world."

IIId. The Free World Must Prevail

Superintelligence will give a decisive economic and military advantage. China isn't at all out of the game yet. In the race to AGI, the free world's very survival will be at stake. Can we maintain our preeminence over the authoritarian powers? And will we manage to avoid self-destruction along the way?

> The story of the human race is War. Except for brief and precarious interludes, there has never been peace in the world; and before history began, murderous strife was universal and unending.
>
> …
>
> Might not a bomb no bigger than an orange be found to possess a secret power to destroy a whole block of buildings — nay, to concentrate the force of a thousand tons of cordite and blast a township at a stroke?
>
> ———————————————————
>
> WINSTON CHURCHILL
> "Shall We All Commit Suicide?" (1924)

SUPERINTELLIGENCE WILL BE the most powerful technology— and most powerful weapon—-mankind has ever developed. It will give a decisive military advantage, perhaps comparable only with nuclear weapons. Authoritarians could use super-intelligence for world conquest, and to enforce total control internally. Rogue states could use it to threaten annihilation. And though many count them out, once the CCP wakes up to AGI it has a clear path to being competitive (at least until and unless we drastically improve US AI lab security).

Every month of lead will matter for safety too. We face the

greatest risks if we are locked in a tight race, democratic allies and authoritarian competitors each racing through the already-precarious intelligence explosion at breakneck pace—forced to throw any caution by the wayside, fearing the other getting superintelligence first. Only if we preserve a healthy lead of democratic allies will we have the margin of error for navigating the extraordinarily volatile and dangerous period around the emergence of superintelligence. And only American leadership is a realistic path to developing a nonproliferation regime to avert the risks of self-destruction superintelligence will unfold.

Our generation too easily takes for granted that we live in peace and freedom. And those who herald the age of AGI in SF too often ignore the elephant in the room: superintelligence is a matter of national security, and the United States must win.

Whoever leads on superintelligence will have a decisive military advantage

Superintelligence is not just any other technology—hypersonic missiles, stealth, and so on—where US and liberal democracies' leadership is highly desirable, but not strictly necessary. The military balance of power can be kept if the US falls behind on one or a couple such technologies; these technologies matter a great deal, but can be outweighed by advantages in other areas.

The advent of superintelligence will put us in a situation unseen since the advent of the atomic era: those who have it will wield complete dominance over those who don't.

I've previously discussed the vast power of superintelligence. It'll mean having billions of automated scientists and engineers and technicians, each much smarter than the smartest human scientists, furiously inventing new technologies, day and night. The acceleration in scientific and technological development will be extraordinary. As superintelligence is applied to R&D in military technology, we could quickly go through decades of military technological progress.

**The Gulf War, or: What a few-decades-worth of techno-
logical lead implies for military power**

The Gulf War provides a helpful illustration of how a 20-
30 year lead in military technology can be decisive. At
the time, Iraq commanded the fourth-largest army in the
world. In terms of numbers (troops, tanks, artillery), the
US-led coalition barely matched (or was outmatched) by
the Iraqis, all while the Iraqis had had ample time to en-
trench their defenses (a situation that would normally
require a 3:1, or 5:1, advantage in military manpower to
dislocate).

But the US-led coalition *obliterated* the Iraqi army in a
merely 100-hour ground war. Coalition dead numbered
a mere 292, compared to 20k-50k Iraqi dead and hundreds
of thousands of others wounded or captured. The Coalition
lost a mere 31 tanks, compared to the destruction of over
3,000 Iraqi tanks.

The difference in technology wasn't godlike or unfath-
omable, but it was utterly and completely decisive: guided
and smart munitions, early versions of stealth, better sen-
sors, better tank scopes (to see farther in the night and in
dust storms), better fighter jets, an advantage in reconnais-
sance, and so on.

(For a more recent example, recall Iran launching a massive
attack of 300 missiles at Israel, "99%" of which were inter-
cepted by superior Israel, US, and allied missile defense.)

A lead of a year or two or three on superintelligence could
mean as *utterly decisive* a military advantage as the US coalition
had against Iraq in the Gulf War. A complete reshaping of the
military balance of power will be on the line.

Imagine if we had gone through the military technological de-
velopments of the 20th century in less than a decade. We'd
have gone from horses and rifles and trenches, to modern tank
armies, in a couple years; to armadas of supersonic fighter
planes and nuclear weapons and ICBMs a couple years after

that; to stealth and precision that can knock out an enemy before they even know you're there another couple years after that.

That is the situation we will face with the advent of superintelligence: the military technological advances of a century compressed to less than a decade. We'll see superhuman hacking that can cripple much of an adversary's military force, roboarmies and autonomous drone swarms, but more importantly completely new paradigms we can't yet begin to imagine, and the inventions of new WMDs with thousandfold increases in destructive power (and new WMD defenses too, like impenetrable missile defense, that rapidly and repeatedly upend deterrence equilibria).

And it wouldn't just be technological progress. As we solve robotics, labor will become fully automated, enabling a broader industrial and economic explosion, too. It is plausible growth rates could go into the 10s of percent a year; within at most a decade, the GDP of those with the lead would trounce those behind. Rapidly multiplying robot factories would mean not only a drastic technological edge, but also production capacity to dominate in pure materiel. Think millions of missile interceptors; billions of drones; and so on.

Of course, we don't know the limits of science and the many frictions that could slow things down. But no godlike advances are necessary for a decisive military advantage. And a billion superintelligent scientists will be able to do *a lot. It seems clear that within a matter of years, pre-superintelligence militaries would become hopelessly outclassed*.

The military advantage would be decisive even against nuclear deterrents

To be even clearer: it seems likely the advantage conferred by superintelligence would be decisive enough even to preemptively take out an adversary's nuclear deterrent. Improved sensor networks and analysis could locate even the quietest current nuclear submarines (similarly for mobile missile launchers). Millions or billions of mouse-sized

autonomous drones, with advances in stealth, could infiltrate behind enemy lines and then surreptitiously locate, sabotage, and decapitate the adversary's nuclear forces. Improved sensors, targeting, and so on could dramatically improve missile defense (similar to, say, the Iran vs. Israel example above); moreover, if there is an industrial explosion, robot factories could churn out thousands of interceptors for each opposing missile. And all of this is without even considering completely new scientific and technological paradigms (e.g., remotely deactivating all the nukes).

It would simply be no contest. And not just no contest in the nuclear sense of "we could mutually destroy each other," but no contest in terms of being able to obliterate the military power of a rival without taking significant casualties. A couple years of lead on superintelligence would mean complete dominance.

If there is a rapid intelligence explosion, it's plausible a lead of mere *months* could be decisive: months could mean the difference between roughly human-level AI systems and substantially superhuman AI systems. Perhaps possessing those initial superintelligences alone, even before being broadly deployed, would be enough for a decisive advantage, e.g. via superhuman hacking abilities that could shut down pre-superintelligence militaries, more limited drone swarms that threaten instant death for every opposing leader, official, and their families, and advanced bioweapons developed with AlphaFold-style simulation that could target specific ethnic groups, e.g. anybody but Han Chinese (or simply withhold the cure from the adversary).

China can be competitive

Many seem complacent about China and AGI. The chip export controls have neutered them, and the leading AI labs are in the US and the UK—so we don't have much to worry about, right? Chinese LLMs are *fine*—they are definitely capable of training large models!—but they are at best comparable to the

second tier of US labs.[99] And even Chinese models are often mere ripoffs of American open source releases (for example, the Yi-34B architecture seems to have essentially the Llama2 architecture, with merely a few lines of code changed[100]). Chinese deep learning used to be more important than it is today (for example Baidu published one of the first modern scaling law papers), and while China publishes more papers in AI than the US, they don't seem to have driven any of the key breakthroughs in recent years.

That's all merely a prelude, however. If and when the CCP wakes up to AGI, we should expect extraordinary efforts on the part of the CCP to compete. And I think there's a pretty clear path for China to be in the game: outbuild the US and steal the algorithms.

1. COMPUTE

1a. Chips: China now seems to have demonstrated the ability to manufacture 7nm chips. While going beyond 7nm will be difficult (requiring EUV), 7nm is enough! For reference, 7nm is what Nvidia A100s used. The indigenous Huawei Ascend 910B, based on the SMIC 7nm platform, seems to only be ~2-3x worse on performance/$ than an equivalent Nvidia chip would be.[101]

The yield of SMIC's 7nm production and the general maturity of Chinese abilities here is debated,[102] and a critical open question is in what quantities they could produce these 7nm chips.[103] Still, it seems like there's at least a very reasonable chance they'll be able to do this at large scale in a few years.

Most of the gains in AI chips have come from improved chip design adapting them for AI use cases (and China likely already steals Nvidia chip designs from the Taiwan supply chain[104]). 7nm vs. 3nm or 2nm, and their general fab immaturity, probably makes things more expensive for China.[105] But that seems by no means fatal; you can make very good AI chips on top of a 7nm process. I wouldn't have high confidence at this point, for example, that they couldn't just spend a bit more and get ample compute for the $100B+ and trillion-dollar

[99] For example, Yi-Large seems to be a GPT-4-class model on LMSys, but that's over a year after OpenAI released GPT-4.

[100] Similarly, Qwen hugging face code cites Mistral a lot, and it seems that Chinese LLM dependence on American open-source is an explicit worry that has made its way to the Chinese Premier.

[101] The Huawei Ascend 910B seems to cost around 120,000 yuan per card, or about $17k. This is produced on the SMIC 7nm node, while performing similarly to an A100. H100s are maybe ~3x better than A100s, while costing somewhat more ($20-25k ASP), suggesting only a ~2-3x cost increase for equivalent AI GPU performance for China right now.

[102] For example, they're still using Western HBM memory (which for some reason is not export controlled?), though CXMT is said to be sampling HBM next year.

[103] Though, since they can still import other types of chips from the West, they could simply direct the entirety of their 7nm node to AI chips, making up for lower overall production.

[104] Notably, even cybercriminals were able to hack Nvidia and get key GPU design secrets. Moreover, TPUv6 designs were apparently among what was stolen by the recently-indicted Chinese national at Google.

[105] For example, maybe these are 2x worse on perf/$ or perf/Watt. In turn, that also means to achieve the same overall datacenter performance, you need more power, and need more chips networked together, which also makes things more of a hassle.

training clusters in a few years.[106]

1b. Outbuilding the US: The binding constraint on the largest training clusters won't be chips, but industrial mobilization—perhaps most of all the 100GW of power for the trillion-dollar cluster. But if there's one thing China can do better than the US it's *building stuff.*

In the last decade, China has roughly built as much new electricity capacity as *the entire US capacity* (while US capacity has remained basically flat). In the US, these things get stuck in environmental review, permitting, and regulation for a decade first. It thus seems quite plausible that China will be able to simply outbuild the US on the largest training clusters.

American and Chinese Power Generation

SITUATIONAL AWARENESS | Leopold Aschenbrenner

Figure 37: The AI power buildout for 2030 seems much more doable for China than the US. Based on earlier estimates from IIIa. Racing to the Trillion-Dollar Cluster
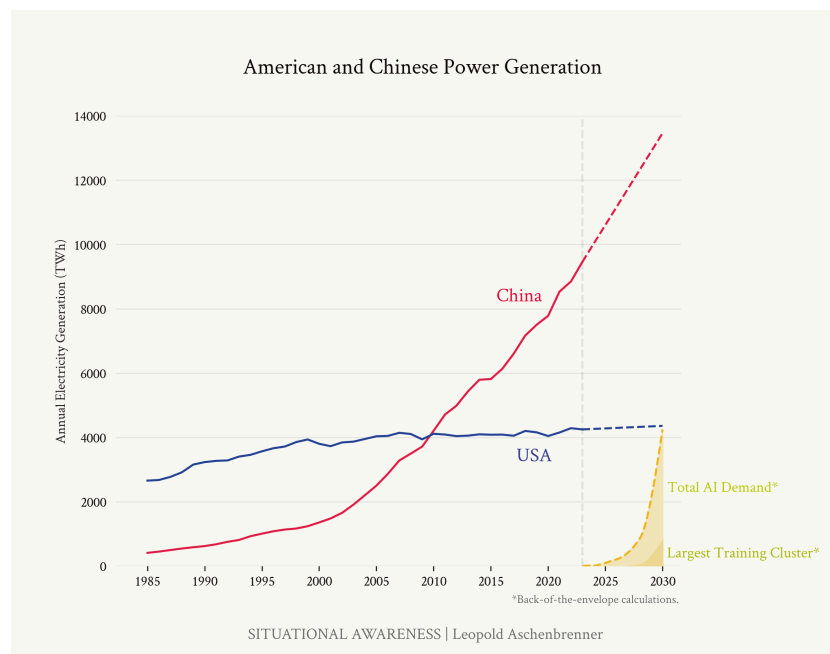
[106] Note that even 3x more on chips would be much less than that in terms of increase in datacenter costs. Actual logic fab costs are <5% of Nvidia GPU cost, and even considering memory and CoWoS it's less than 20% of the Nvidia pricetag due to their margin. And even GPUs themselves tend to be only 50-60% of the cost of a datacenter. So 3x more expensive on the chip fab end might translate into much, much less than a 3x increase in overall cost. Even for 10x more expensive chips, it seems like China could stomach that without hugely increasing datacenter costs.

2. ALGORITHMS

As discussed extensively in Counting the OOMs, scaling compute is only part of the story: algorithmic advances probably contribute at least half of AI progress. We're developing the key algorithmic breakthroughs for AGI *right now* (essentially the EUV of algorithms because of the data wall).

By default, I expect Western labs to be well ahead; they have much of the key talent, and in recent years have developed all of the key breakthroughs. The size of the advantage may well be equivalent to a 10x (or even 100x) bigger cluster in a few years; this would provide the United States with a reasonably comfortable lead.

And yet, on the current course, we will completely surrender this advantage: as discussed extensively in the security section, the current state of security essentially makes it trivial for China to infiltrate American labs. And so, unless we lock down the labs very soon, I expect China to be able to simply steal the key algorithmic ingredients for AGI, and match US capabilities.[107]

(Even worse, if we don't improve security, there is an even more salient path for China to compete. They won't even need to train their own AGI: they'll just be able to steal the AGI weights directly. Once they've stolen a copy of the automated AI researcher, they'll be off to the races, and can launch their own intelligence explosion. If they're willing to apply less caution—both good caution, and unreasonable regulation and delay—than the US, they could race through the intelligence explosion more quickly, outrunning us to superintelligence.)

To DATE, US tech companies have made a much bigger bet on AI and scaling than any Chinese efforts; consequently, we are well ahead. But counting out China now is a bit like counting out Google in the AI race when ChatGPT came out in late 2022. Google hadn't yet focused their efforts in an intense AI bet, and it looked as though OpenAI was far ahead—but once Google woke up, a year and half later, they are putting up a very serious fight. China, too, has a clear path to putting up a very serious fight. If and when the CCP mobilizes in the race to AGI, the picture could start looking very different.

Perhaps the Chinese government will be incompetent; perhaps they decide AI threatens the CCP and impose stifling regulation. But I wouldn't count on it.

[107] Some argue that even if China stole these secrets, they wouldn't be able to compete because it requires tacit knowledge. I disagree. I think of this as having two layers. The bottom layer is the engineering prowess for large-scale training runs; these training runs can be hacky and delicate and requires tacit knowledge. But as I'll discuss later, Chinese AI efforts have shown themselves perfectly capable of training large-scale models, and I think they will have this tacit knowledge indigenously. The top layer is the algorithmic recipe—things like model architecture, the right scaling laws, etc.—that could be conveyed in a one-hour call. These compute multipliers are usually discrete changes, meaning the underlying tacit knowledge for large-scale training runs should transfer. I don't think "tacit knowledge" will be a decisive barrier for Chinese AGI efforts.

I, for one, think we need to operate under the assumption that we will face a full-throated Chinese AGI effort. As every year we get dramatic leaps in AI capability, as we start seeing early automation of software engineers, as AI revenue explodes and we start seeing $10T valuations and trillion-dollar cluster build-outs, as a broader consensus starts to form that we are on the cusp of AGI—the CCP will take note. Much as I expect these leaps to wake up the USG to AGI, I would expect it to wake up the CCP to AGI—and to wake up to what being behind on AGI would mean for their national power.

They will be a formidable adversary.

The authoritarian peril

A dictator who wields the power of superintelligence would command concentrated power unlike any we've ever seen. In addition to being able to impose their will on other countries, they could enshrine their rule internally. Millions of AI-controlled robotic law enforcement agents could police their populace; mass surveillance would be hypercharged; dictator-loyal AIs could individually assess every citizen for dissent, with advanced near-perfect lie detection rooting out any disloyalty.

Most importantly, the robotic military and police force could be wholly controlled by a single political leader, and programmed to be perfectly obedient—no more risk of coups or popular rebellions.

Whereas past dictatorships were never permanent, superintelligence could eliminate basically all historical threats to a dictator's rule and lock in their power (cf value lock-in). If the CCP gets this power, they could enforce the Party's conception of "truth" totally and completely.

To be clear, I don't just worry about dictators getting superintelligence because "our values are better." I believe in freedom and democracy, strongly, because I don't know what the right values are. In the long arc of history, "time has upset many

fighting faiths." I believe we should place our faith in mech-
anisms of error correction, experimentation, competition, and
adaption.

Superintelligence will give those who wield it the power to
crush opposition, dissent, and lock in their grand plan for hu-
manity. It will be difficult for anyone to resist the terrible temp-
tation to use this power. I hope, dearly, that we can instead
rely on the wisdom of the Framers—letting radically different
values flourish, and preserving the raucous plurality that has
defined the American experiment.

At stake in the AGI race will not just be the advantage in some
far-flung proxy war, but whether freedom and democracy can
survive for the next century and beyond. The course of hu-
man history is as brutal as it is clear. Twice in the 20th century
tyranny threatened the globe; we must be under no delusion
that this threat is banished forever. For many of my young
friends, freedom and democracy feel like a given—but they
are not. By far the most common political system in history is
authoritarianism.[108]

[108] Primarily monarchy.

I genuinely do not know the intentions of the CCP and their
authoritarian allies. But, as a reminder: the CCP is a regime
founded on the continued worship of perhaps the greatest to-
talitarian mass-murderer in human history ("with estimates
ranging from 40 to 80 million victims due to starvation, per-
secution, prison labor, and mass executions"); a regime that
recently put a million Uyghurs in concentration camps and
crushed a free Hong Kong; a regime that systematically prac-
tices mass surveillance for social control, both of the new-
fangled (tracking phones, DNA databases, facial recognition,
and so on) and the old-fangled (recruiting an army of citizens
to report on their neighbors) kind; a regime that ensures all
text messages pass through a censor, and that goes so far to re-
press dissent as to pull families into police stations when their
child overseas attends a protest; a regime that has cemented
Xi Jinping as dictator-for-life; a regime that touts its aims to
militarily crush and "reeducate" a free neighboring nation; a
regime that explicitly seeks a China-centric world order.

*The free world must prevail over the authoritarian powers in this
race.* We owe our peace and freedom to American economic
and military preeminence. Perhaps even empowered with
superintelligence, the CCP will behave responsibly on the in-
ternational stage, leaving each to their own. But the history of
dictators of their ilk is not pretty. If America and her allies fail
to win this race, we risk it all.

Maintaining a healthy lead will be decisive for safety

It is the cursed history of science and technology that as they
have unfolded their wonders, they have also expanded the
means of destruction: from sticks and stones, to swords and
spears, rifles and cannons, machine guns and tanks, bombers
and missiles, nuclear weapons. The "destruction/$" curve
has consistently gone down as technology has advanced. We
should expect the rapid technological progress post-superintelligence
to follow this trend.

Perhaps dramatic advances in biology will yield extraordi-
nary new bioweapons, ones that spread silently, swiftly, be-
fore killing with perfect lethality on command (and that can
be made extraordinarily cheaply, affordable even for terrorist
groups). Perhaps new kinds of nuclear weapons enable the
size of nuclear arsenals to increase by orders of magnitude,
with new delivery mechanisms that are undetectable. Perhaps
mosquito-sized drones, each carrying a deadly poison, could
be targeted to kill every member of an opposing nation. It's
hard to know what a century's worth of technological progress
would yield—but I am confident it would unfold appalling
possibilities.

Humanity barely evaded self-destruction during the Cold War.
On the historical view, the greatest existential risk posed by
AGI is that it will enable us to develop extraordinary new
means of mass death. This time, these means could even pro-
liferate to become accessible to rogue actors or terrorists (espe-
cially if, as on the current course, the superintelligence weights
aren't sufficiently protected, and can be directly stolen by
North Korea, Iran, and co.).

North Korea already has a concerted bioweapons program: the US assesses that "North Korea has a dedicated, national level offensive program" to develop and produce bioweapons. It seems plausible that their primary constraint is how far their small circle of top scientists has been able to push the limits of (synthetic) biology. What happens when that constraint is removed, when they can use millions of superintelligences to accelerate their bioweapons R&D? For example, the US assesses that North Korea currently has "limited ability" to genetically engineer biological products—what happens when that becomes unlimited? With what unholy new concoctions will they hold us hostage?

Moreover, as discussed in the superalignment section, there will be extreme safety risks around and during the intelligence explosion—we will be faced with novel technical challenges to ensure we can reliably trust and control superhuman AI systems. This very well may require us to slow down at some critical moments, say, delaying by 6 months in the middle of the intelligence explosion to get additional assurances on safety, or using a large fraction of compute on alignment research rather than capabilities progress.

Some hope for some sort of international treaty on safety. This seems fanciful to me. The world where both the CCP and USG are AGI-pilled enough to take safety risk seriously is also the world in which both realize that international economic and military predominance is at stake, that being months behind on AGI could mean being permanently left behind. If the race is tight, any arms control equilibrium, at least in the early phase around superintelligence, seems extremely unstable. In short, "breakout" is too easy: the incentive (and the fear that others will act on this incentive) to race ahead with an intelligence explosion, to reach superintelligence and the decisive advantage, too great.[109] At the very least, the odds we get something good-enough here seem slim. (How have those climate treaties gone? That seems like a dramatically easier problem compared to this.)

The main—perhaps the only—hope we have is that an alliance of democracies has a healthy lead over adversarial powers. The

[109] Consider the following comparison of unstable vs. stable arms control equilibria. 1980s arms control during the Cold War reduced nuclear weapons substantially, but targeted a stable equilibrium. The US and the Soviet Union still had 1000s of nuclear weapons. MAD was assured, even if one of the actors tried a crash program to build more nukes; and a rogue nation could try to build some nukes of their own, but not fundamentally threaten the US or Soviet Union with overmatch.

However, when disarmament is to very low levels of weapons or occurs amidst rapid technological change, the equilibrium is unstable. A rogue actor or treaty-breaker can easily start a crash program and threaten to totally overmatch the other players. Zero nukes wouldn't be a stable equilibrium; similarly, this paper has interesting historical case studies (such as post-WWI arms limitations and the Washington Naval Treaty) where disarmament in similarly dynamic situations destabilized, rather than stabilized.

If mere months of lead on AGI would give an utterly decisive advantage, this makes stable disarmament on AI similarly difficult. A rogue upstart or a treaty-breaker could gain a huge edge by secretly starting a crash program; the temptation would be too great for any sort of arrangement to be stable.

United States must lead, and use that lead to enforce safety norms on the rest of the world. That's the path we took with nukes, offering assistance on the peaceful uses of nuclear technology in exchange for an international nonproliferation regime (ultimately underwritten by American military power)—and it's the only path that's been shown to work.

Perhaps most importantly, a healthy lead gives us room to maneuver: the ability to "cash in" parts of the lead, if necessary, to get safety right, for example by devoting extra work to alignment during the intelligence explosion.

The safety challenges of superintelligence would become *extremely* difficult to manage if you are in a neck-and-neck arms race. A 2 year vs. a 2 month lead could easily make all the difference. If we have only a 2 month lead, we have no margin at all for safety. In fear of the CCP's intelligence explosion, we'd almost certainly race, no holds barred, through our own intelligence explosion—barreling towards AI systems vastly smarter than humans in months, without any ability to slow down to get key decisions right, with all the risks of superintelligence going awry that implies. We'd face an extremely volatile situation, as we and the CCP rapidly developed extraordinary new military technology that repeatedly destabilized deterrence. If our secrets and weights aren't locked down, it might even mean a range of other rogue states are close as well, each of them using superintelligence to furnish their own new arsenal of super-WMDs. Even if we barely managed to inch out ahead, it would likely be a pyrrhic victory; the existential struggle would have brought the world to the brink of total self-destruction.

Superintelligence looks very different if the democratic allies have a healthy lead, say 2 years.[110] That buys us the time necessary to navigate the unprecedented series of challenges we'll face around and after superintelligence, and to stabilize the situation.

If and when it becomes clear that the US will decisively win, *that's* when we offer a deal to China and other adversaries. They'll know they won't win, and so they'll know their only

[110] Note that, given the already-rapid pace of AI progress today, and the even-more-rapid pace we should expect in an intelligence explosion, and the broader technological explosion post-superintelligence, "even" a 2-year-lead would be a *huge* difference in capabilities in the post-superintelligence world.

option is to come to the table; and we'd rather avoid a fever-ish standoff or last-ditch military attempts on their part to sabotage Western efforts. In exchange for guaranteeing non-interference in their affairs, and sharing the peaceful benefits of superintelligence, a regime of nonproliferation, safety norms, and a semblance of stability post-superintelligence can be born.

In any case, as we go deeper into this struggle, we must not forget the threat of self-destruction. That we made it through the Cold War in one piece involved too much luck[111]—and the destruction could be a thousandfold more potent than what we faced then. A healthy lead by an American-led coalition of democracies—and a solemn exercise of this leadership to stabilize whatever volatile situation we find ourselves in—is probably the safest path to navigating past this precipice. But in the heat of the AGI race, we better not screw it up.

[111] Daniel Ellsberg recounts this riv-etingly, as one of the nuclear war planners at RAND and in the national security apparatus at the time.

Superintelligence is a matter of national security

It is clear: AGI is an existential challenge for the national secu-rity of the United States. It's time to start treating it as such.

Slowly, the USG is starting to move. The export controls on American chips are a huge deal, and were an incredibly pre-scient move at the time. But we have to get serious across the board.

The US *has* a lead. We just have to keep it. And we're screwing that up right now. Most of all, we must rapidly and radically lock down the AI labs, before we leak key AGI breakthroughs in the next 12-24 months (or the AGI weights themselves). We must build the compute clusters in the US, not in dictator-ships that offer easy money. And yes, American AI labs have a duty to work with the intelligence community and the military. America's lead on AGI won't secure peace and freedom by just building the best AI girlfriend apps. It's not pretty—but we must build AI for American defense.

WE ARE ALREADY on course for the most combustive international situation in decades. Putin is on the march in Eastern Europe. The Middle East is on fire. The CCP views taking Taiwan as its destiny. Now add in the race to AGI. Add in a century's worth of technological breakthroughs compressed into years post-superintelligence. It will be the one of most unstable international situations ever seen—and at least initially, the incentives for first-strikes will be tremendous.[112]

There's already an eerie convergence of AGI timelines (~2027?) and Taiwan watchers' Taiwan invasion timelines (China ready to invade Taiwan by 2027?)—a convergence that will surely only heighten as the world wakes up to AGI. (Imagine if in 1960, the vast majority of the world's uranium deposits were somehow concentrated in Berlin!) It seems to me that there is a real chance that the AGI endgame plays out with the backdrop of world war. Then all bets are off.

[112] Everyone will be racing to their own superintelligences, and there will be a limited window before someone ahead will have irreversibly pulled away. There will be a big incentive to try to disable the enemy superintelligence clusters before they've gained a sufficient physical advantage (e.g., using superintelligence to develop impenetrable missile defense or drone swarms) that leaves everyone else permanently in the dust.

IV. The Project

As the race to AGI intensifies, the national security state will get involved. The USG will wake from its slumber, and by 27/28 we'll get some form of government AGI project. No startup can handle superintelligence. Somewhere in a SCIF, the endgame will be on.

> We must be curious to learn how such a set of objects—hundreds of power plants, thousands of bombs, tens of thousands of people massed in national establishments—can be traced back to a few people sitting at laboratory benches discussing the peculiar behavior of one type of atom.
>
> SPENCER R. WEART

MANY PLANS FOR "AI governance" are put forth these days, from licensing frontier AI systems to safety standards to a public cloud with a few hundred million in compute for academics. These seem well-intentioned—but to me, it seems like they are making a category error.

*I find it an **insane** proposition that the US government will let a random SF startup develop superintelligence.* Imagine if we had developed atomic bombs by letting Uber just improvise.

Superintelligence—AI systems *much* smarter than humans— will have vast power, from developing novel weaponry to driving an explosion in economic growth. Superintelligence will be

the locus of international competition; a lead of months potentially decisive in military conflict.

It is a delusion of those who have unconsciously internalized our brief respite from history that this will not summon more primordial forces. Like many scientists before us, the great minds of San Francisco hope that they can control the destiny of the demon they are birthing. Right now, they still can; for they are among the few with situational awareness, who understand what they are building. But in the next few years, the world will wake up. So too will the national security state. History will make a triumphant return.

As in many times before—Covid, WWII—it will seem as though the United States is asleep at the wheel—before, all at once, the government shifts into gear in the most extraordinary fashion. There will be a moment—in just a few years, just a couple more "2023-level" leaps in model capabilities and AI discourse—where it will be clear: we are on the cusp of AGI, and superintelligence shortly thereafter. While there's a lot of flux in the exact mechanics, one way or another, the USG will be at the helm; the leading labs will ("voluntarily") merge; Congress will appropriate trillions for chips and power; a coalition of democracies formed.

Startups are great for many things—but a startup on its own is simply not equipped for being in charge of the United States' most important national defense project. We will need government involvement to have even a hope of defending against the all-out espionage threat we will face; the private AI efforts might as well be directly delivering superintelligence to the CCP. We will need the government to ensure even a semblance of a sane chain of command; you can't have random CEOs (or random nonprofit boards) with the nuclear button. We will need the government to manage the severe safety challenges of superintelligence, to manage the fog of war of the intelligence explosion. We will need the government to deploy superintelligence to defend against whatever extreme threats unfold, to make it through the extraordinarily volatile and destabilized international situation that will follow. We will need the government to mobilize a democratic coalition to win the race with

authoritarian powers, and forge (and enforce) a nonproliferation regime for the rest of the world. I wish it weren't this way—but we will need the government. (Yes, regardless of the Administration.)

In any case, my main claim is not normative, but descriptive. In a few years, The Project will be on.

The path to The Project

A turn-of-events seared into my memory is late February to mid-March of 2020. In those last weeks of February and early days of March, I was in utter despair: it seemed clear that we were on the covid-exponential: a plague was about to sweep the country, the collapse of our hospitals was imminent—and yet almost nobody took it seriously. The Mayor of New York was still dismissing Covid-fears as racism and encouraging people to go to Broadway shows. All I could do was buy masks and short the market.

And yet within just a few weeks, the entire country shut down and Congress had appropriated trillions of dollars (literally >10% of GDP). Seeing where the exponential might go ahead of time was too hard, but when the threat got close enough, existential enough, extraordinary forces were unleashed. The response was late, crude, blunt—but it came, and it was dramatic.

The next few years in AI will feel similar. We're in the midgame now. 2023 was already a wild shift. AGI went from a fringe topic you'd be hesitant to associate with, to the subject of major Senate hearings and summits of world leaders. Given how early we are still, the level of USG engagement has been impressive to me. A couple more "2023"s, and the Overton window will be blown completely open.

As we race through the OOMs, the leaps will continue. By 2025/2026 or so I expect the next truly shocking step-changes; AI will drive $100B+ annual revenues for big tech companies and outcompete PhDs in raw problem-solving smarts. Much as

the Covid stock-market collapse made many take Covid seri-
ously, we'll have $10T companies and the AI mania will be ev-
erywhere. If that's not enough, by 2027/28, we'll have models
trained on the $100B+ cluster; full-fledged AI agents/drop-in
remote workers will start to widely automate software engi-
neering and other cognitive jobs. Each year, the acceleration
will feel dizzying.

While many don't yet see the possibility of AGI, eventually
a consensus will form. Some, like Szilard, saw the possibility
of an atomic bomb much earlier than others. Their alarm was
not well-received initially; the possibility of a bomb was dis-
missed as remote (or at least, it was felt that the conservative
and proper thing was to play down the possibility). Szilard's
fervent secrecy appeals were mocked and ignored. But many
scientists, initially skeptical, started realizing a bomb was possi-
ble as more and more empirical results came in. Once a major-
ity of scientists came to believe we were on the cusp of a bomb,
the government, in turn, saw the national security exigency as
too great—and the Manhattan Project got underway.

As the OOMs go from theoretical extrapolation to (extraordi-
nary) empirical reality, gradually, a consensus will form, too,
among the leading scientists and executives and government
officials: we are on the cusp, on the cusp of AGI, on the cusp
of an intelligence explosion, on the cusp of superintelligence.
And somewhere along here, we'll get the first genuinely terri-
fying demonstrations of AI: perhaps the oft-discussed "helping
novices make bioweapons," or autonomously hacking critical
systems, or something else entirely. It will become clear: this
technology will be an utterly decisive military technology. Even
if we're lucky enough to not be in a major war, it seems likely
that the CCP will have taken notice and launched a formidable
AGI effort. Perhaps the eventual (inevitable) discovery of the
CCP's infiltration of America's leading AI labs will cause a big
stir.

Somewhere around 26/27 or so, the mood in Washington will
become somber. People will start to viscerally feel what is hap-
pening; they will be *scared*. From the halls of the Pentagon to
the backroom Congressional briefings will ring the obvious

question, the question on everybody's minds: do we need an
AGI Manhattan Project? Slowly at first, then all at once, it will
become clear: this is happening, things are going to get *wild*,
this is the most important challenge for the national security
of the United States since the invention of the atomic bomb. In
one form or another, the national security state will get very
heavily involved. The Project will be the necessary, indeed the
only plausible, response.

Of course, this is an extremely abbreviated account—a lot de-
pends on when and how consensus forms, key warning shots,
and so on. DC is infamously dysfunctional. As with Covid,
and even the Manhattan Project, the government will be incred-
ibly late and hamfisted. After Einstein's letter to the President
in 1939 (drafted by Szilard), an Advisory Committee on Ura-
nium was formed. But officials were incompetent, and not
much happened initially. For example, Fermi only got $6k
(about $135k in today's dollars) to support his research, and
even that was not given easily and only received after months
of waiting. Szilard believed that the project was delayed for
at least a year by the short-sightedness and sluggishness of
the authorities. In March 1941, the British government finally
concluded a bomb was inevitable. The US committee initially
entirely ignored this British report for months—until finally in
December 1941, a full-scale atomic bomb effort was launched.

There are many ways this could be operationalized in practice.
To be clear, this doesn't need to look like literal nationaliza-
tion, with AI lab researchers now employed by the military or
whatever (though it might!).[113] Rather, I expect a more suave
orchestration. The relationship with the DoD might look like
the relationship the DoD has with Boeing or Lockheed Mar-
tin. Perhaps via defense contracting or similar, a joint venture
between the major cloud compute providers, AI labs, and the
government is established, making it functionally a project of
the national security state. Much like the AI labs "voluntarily"
made commitments to the White House in 2023, Western labs
might more-or-less "voluntarily" agree to merge in the national
effort. And likely Congress will have to be involved, given the
trillions of investment involved, and for checks-and-balances.[114]
How all these details shake out is a story for another day.

[113] Note that while private companies
help develop components for nuclear
weapons, they are never allowed to
possess a completed and assembled
nuclear weapon. In comparison, the
mainline version of the "AGI govern-
ment project" I am putting forward
here is unprecedentedly privatized, for
the WMD reference class.

[114] Congress—even the Vice President!—
didn't know about the Manhattan
Project. We probably shouldn't repeat
that here; I'd even suggest that key
officials for The Project require Senate
confirmation.

But by late 26/27/28 it will be underway. The core AGI re-
search team (a few hundred researchers) will move to a secure
location; the trillion-dollar cluster will be built in record-speed;
The Project will be on.

Why The Project is the only way

I AM UNDER no illusions about the government. Governments
face all sorts of limitations and poor incentives. I am a big be-
liever in the American private sector, and would almost never
advocate for heavy government involvement in technology or
industry.

I used to apply this same framework to AGI—until I joined an
AI lab. AI labs are very good at some things: they've been able
to take AI from an academic science project to the commercial
big stage, in a way only a startup can. But ultimately, AI labs
are still startups. We simply shouldn't expect startups to be
equipped to handle superintelligence.

There are no good options here—but I don't see another way.
When a technology becomes this important for national secu-
rity, we will need the USG.

*Superintelligence will be the United States' most important national de-
fense project*

I've discussed the power of superintelligence in previous
pieces. Within years, superintelligence would completely shake
up the military balance of power. By the early 2030s, the en-
tirety of the US arsenal (like it or not, the bedrock of global
peace and security) will probably be obsolete. It will not just be
a matter of modernization, but a wholesale replacement.

Simply put, it will become clear that the development of AGI
will fall in a category more like nukes than the internet. Yes,
of course it'll be dual-use—but nuclear technology was dual-
use too. The civilian applications will have their time. But in

the fog of the AGI endgame, for better or for worse, national security will be the primary backdrop.

We will need to completely reshape US forces, within a matter of years, in the face of rapid technological change—or risk being completely outmatched by adversaries who do. Perhaps most of all, the initial priority will be to deploy superintelligence for defensive applications, to develop countermeasures to survive untold new threats: adversaries with superhuman hacking capabilities, new classes of stealthy drone swarms that could execute a preemptive strike on our nuclear deterrent, the proliferation of advances in synthetic biology that can be weaponized, turbulent international (and national) power struggles, and rogue superintelligence projects.

Whether nominally private or not, the AGI project will need to be, *will be,* integrally a defense project, and it will require extremely close cooperation with the national security state.

A sane chain of command for superintelligence

The power—and the challenges—of superintelligence will fall into a very different reference class than anything else we're used to seeing from tech companies. It seems pretty clear: this should not be under the unilateral command of a random CEO. Indeed, in the private-labs-developing-superintelligence world, it's quite plausible individual CEOs would have the power *to literally coup the US government.*[115] Imagine if Elon Musk had final command of the nuclear arsenal.[116] (Or if a random nonprofit board could decide to seize control of the nuclear arsenal.)

It is perhaps obvious, but: as a society, we've decided democratic governments should control the military;[117] superintelligence will be, at least at first, the most powerful military weapon. The radical proposal is not The Project; the radical proposal is taking a bet on private AI CEOs wielding military power and becoming benevolent dictators.

(Indeed, in the private AI lab world, it would likely be even worse than random CEOs with the nuclear button—part of AI

[115] It wouldn't even require cooperation from AI lab employees at this point, since they'll have been mostly automated by this point.

[116] And as Sam Altman once said, every year we get closer to AGI everybody will gain +10 crazy points.

[117] In fact, the government having the biggest guns was an enormous civilizational achievement! Rather than medieval-like fights of all against all, we would sort out disagreements via courts, pluralistic institutions, and so on.

labs' abysmal security is their utter lack of internal controls. That is, random AI lab employees (with zero vetting) could go rogue unnoticed.)

We will need a sane chain of command—along with all the other processes and safeguards that necessarily come with responsibly wielding what will be comparable to a WMD—and it'll require the government to do so. In some sense, this is simply a Burkean argument: the institutions, constitutions, laws, courts, checks and balances, norms and common dedication to the liberal democratic order (e.g., generals refusing to follow illegal orders), and so on that check the power of the government have withstood the test of hundreds of years. Special AI lab governance structures, meanwhile, collapsed the first time they were tested. The US military could already kill basically every civilian in the United States, or seize power, if it wanted to—and the way we keep government power over nuclear weapons in check is not through lots of private companies with their own nuclear arsenals. There's only one chain of command and set of institutions that has proven itself up to this task.

Again, perhaps you are a true libertarian and disagree normatively (let Elon Musk and Sam Altman command their own nuclear arsenals!).[118] But once it becomes clear that superintelligence is a principal matter of national security, I'm sure this is how the men and women in DC will look at it.

The civilian uses of superintelligence

Of course, that doesn't mean the civilian applications of superintelligence will be reserved for the government.

- The nuclear chain reaction was first harnessed as a government project—and nuclear weapons permanently reserved for the government—but civilian nuclear energy flourished as private projects (in the 60s and 70s, before environmentalists shut it down).

- Boeing made the B-29 (the most expensive defense R&D project during WWII, more expensive than the Manhat-

[118] Or perhaps you say, just open-source everything. The issue with simply open sourcing everything is that it's not a happy world of a thousand flowers blooming in the US, but a world in which the CCP has free access to US-developed superintelligence, and can outbuild (and apply less caution/regulation) and take over the world. And the other issue, of course, is the proliferation of super-WMDs to every rogue state and terrorist group in the world. I don't think it'll end well. It's a bit like how having no government at all is more likely to lead to tyranny (or destruction) than freedom.

In any case, people overrate the importance of open-source as we get closer to AGI. Given cluster costs escalating to hundreds of billions, and key algorithmic secrets now being proprietary rather than published as they were a couple years ago, it'll be 2-3 or so leading players, rather than some happy community of decentralized coders building AGI.

I do think a different variant of open source will continue to play an important role: models that lag a couple years behind being open sourced, helping the benefits of the technology diffuse broadly.

tan Project) and the B-47 and B-52 long-range bombers in partnership with the military—before using that technology for the Boeing 707, the commercial plane that ushered in the jet era. And today, while Boeing can only sell stealth fighter jets to the government, it can freely develop and sell civilian jets privately.

• And so it went for radar, satellites, rockets, gene technology, WWII factories, and so on.

The initial development of superintelligence will be dominated by the national security exigency to survive and stabilize an incredibly volatile period. And the military uses of superintelligence will remain reserved for the government, and safety norms will be enforced. But once the initial peril has passed, the natural path is for the companies involved in the national consortium (and others) to privately pursue civilian applications.

Even in worlds with The Project, a private, pluralistic, market-based, flourishing ecosystem of civilian applications of superintelligence will have its day.

Security

I've gone on about this at length in a previous piece in the series. On the current course, we may as well give up on having any American AGI effort; China can promptly steal all the algorithmic breakthroughs and the model weights (literally a copy of superintelligence) directly. It's not even clear we'll get to "North Korea-proof" security for superintelligence on the current course. In the private-startups-developing-AGI-world, superintelligence would proliferate to dozens of rogue states. It's simply untenable.

If we're going to be at all serious about this, we obviously need to lock this stuff down. Most private companies have failed to take this seriously. But in any case, if we are to eventually face the full force of Chinese espionage (e.g., stealing the weights being the MSS's #1 priority), it's probably impossible for a private company to get good enough security. It will require

extensive cooperation from the US intelligence community at that point to sufficiently secure AGI. This will involve invasive restrictions on AI labs and on the core team of AGI researchers, from extreme vetting to constant monitoring to working from a SCIF to reduced freedom to leave; and it will require infrastructure only the government can provide, ultimately including the physical security of the AGI datacenters themselves.

In some sense, security alone is sufficient to necessitate the government project—both the free world's preeminence and AI safety are doomed if we can't lock this stuff down. (In fact, I think it's fairly likely to be a major factor in the ultimate trigger: once the Chinese infiltration of the AGI labs becomes clear, every Senator and Congressperson and national security official will... have a strong opinion on the matter.)

Safety

Simply put: there are a lot of ways for us to mess this up— from ensuring we can reliably control and trust the billions of superintelligent agents that will soon be in charge of our economy and military (the superalignment problem), to controlling the risks of misuse of new means of mass destruction.

Some AI labs claim to be committed to safety: acknowledging that what they are building, if gone awry, could cause catastrophe and promising that they will do what is necessary when the time comes. I do not know if we can trust their promise enough to stake the lives of every American on it. More importantly, so far, they have not demonstrated the competence, trustworthiness, or seriousness necessary for what they themselves acknowledge they are building.

At core, they are startups, with all the usual commercial incentives. Competition could push all of them to simply race through the intelligence explosion, and there will at least be some actors that will be willing to throw safety by the wayside. In particular, we may want to "spend some of our lead" to have time to solve safety challenges, but Western labs will need to coordinate to do so. (And of course, private labs will have already had their AGI weights stolen, so their safety pre-

cautions won't even matter; we'll be at the mercy of the CCP's and North Korea's safety precautions.)

One answer is regulation. That may be appropriate in worlds in which AI develops more slowly, but I fear that regulation simply won't be up to the nature of the challenge of the intelligence explosion. What's necessary will be less like spending a few years doing careful evaluations and pushing some safety standards through a bureaucracy. *It'll be more like fighting a war*.

We'll face an insane year in which the situation is shifting extremely rapidly every week, in which hard calls based on ambiguous data will be life-or-death, in which the solutions—even the problems themselves—won't be close to fully clear ahead of time but come down to competence in a "fog of war," which will involve insane tradeoffs like "some of our alignment measurements are looking ambiguous, we don't really understand what's going on anymore, it might be fine but there's some warning signs that the next generation of superintelligence might go awry, should we delay the next training run by 3 months to get more confidence on safety—but oh no, the latest intelligence reports indicate China stole our weights and is racing ahead on their own intelligence explosion, what should we do?".

I'm not confident that a government project would be competent in dealing with this—but the "superintelligence developed by startups" alternative seems much closer to "praying for the best" than commonly recognized. We'll need a chain of command that can bring to the table the seriousness that making these difficult tradeoffs will require.

Stabilizing the international situation

The intelligence explosion and its immediate aftermath will bring forth one of the most volatile and tense situations mankind has ever faced. Our generation is not used to this. But in this initial period, the task at hand will not be to build cool products. It will be to somehow, desperately, make it through this period.

We'll need the government project to win the race against the

authoritarian powers—and to give us the clear lead and breath-ing room necessary to navigate the perils of this situation. We might as well give up if we can't prevent the instant theft of superintelligence model weights. We will want to bundle West-ern efforts: bring together our best scientists, use every GPU we can find, and ensure the trillions of dollars of cluster build-outs happen in the United States. We will need to protect the datacenters against adversary sabotage, or outright attack.

Perhaps, most of all, it will take American leadership to develop—and if necessary, enforce—a nonproliferation regime. We'll need to subvert Russia, North Korea, Iran, and terrorist groups from using their own superintelligence to develop technology and weaponry that would let them hold the world hostage. We'll need to use superintelligence to harden the security of our critical infrastructure, military, and government to defend against extreme new hacking capabilities. We'll need to use superintelligence to stabilize the offense/defense balance of advances in biology or similar. We'll need to develop tools to safely control superintelligence, and to shut down rogue su-perintelligences that come out of others' uncareful projects.AI systems and robots will be moving at 10-100x+ human speed; everything will start happening extremely quickly. We need to be ready to handle whatever other six-sigma upheavals—and concomitant threats—come out of compressing a century's worth of technological progress into a few years.

At least in this initial period, we will be faced with the most extraordinary national security exigency. Perhaps, nobody is up for this task. But of the options we have, The Project is the only sane one.

The Project is inevitable; whether it's good is not

ULTIMATELY, MY MAIN CLAIM here is descriptive: whether we like it or not, superintelligence won't look like an SF startup, and in some way will be primarily in the domain of national security. I've brought up The Project a lot to my San Francisco friends in the past year. Perhaps what's surprised me most is

how surprised most people are about the idea. They simply haven't considered the possibility. But once they consider it, most agree that it seems obvious. If we are at all right about what we think we are building, *of course,* by the end this will be (in some form) a government project. If a lab developed literal superintelligence tomorrow, *of course* the Feds would step in.

One important free variable is not *if* but *when.* Does the government not realize what's happening until we're in the middle of an intelligence explosion—or will it realize a couple years beforehand? If the government project is inevitable, earlier seems better. We'll dearly need those couple years to do the security crash program, to get the key officials up to speed and prepared, to build a functioning merged lab, and so on. It'll be far more chaotic if the government only steps in at the very end (and the secrets and weights will have already been stolen).

Another important free variable is the international coalition we can rally: both a tighter alliance of democracies for developing superintelligence, and a broader benefit-sharing offer made to the rest of the world.

- The former might look like the Quebec Agreement: a secret pact between Churchill and Roosevelt to pool their resources to develop nuclear weapons, while not using them against each other or against others without mutual consent. We'll want to bring in the UK (Deepmind), East Asian allies like Japan and South Korea (chip supply chain), and NATO/other core democratic allies (broader industrial base). A united effort will have more resources, talent, and control the whole supply chain; enable close coordination on safety, national security, and military challenges; and provide helpful checks and balances on wielding the power of superintelligence.

- The latter might look like Atoms for Peace, the IAEA, and the NPT. We should offer to share the peaceful benefits of superintelligence with a broader group of countries (including non-democracies), and commit to not offensively using superintelligence against them. In exchange, they refrain from pursuing their own superintelligence projects, make

safety commitments on the deployment of AI systems, and accept restrictions on dual-use applications. The hope is that this offer reduces the incentives for arms races and proliferation, and brings a broad coalition under a US-led umbrella for the post-superintelligence world order.

Perhaps the most important free variable is simply whether the inevitable government project will be competent. How will it be organized? How can we get this done? How will the checks and balances work, and what does a sane chain of command look like? Scarcely any attention has gone into figuring this out.[119] Almost all other AI lab and AI governance politicking is a sideshow. This is the ballgame.

[119] To my Progress Studies brethren: you should think about this, this will be the culmination of your intellectual project! You spend all this time studying American government research institutions, their decline over the last half-century, and what it would take to make them effective again. Tell me: how will we make The Project effective?

The endgame

AND SO BY 27/28, the endgame will be on. By 28/29 the intelligence explosion will be underway; by 2030, we will have summoned superintelligence, in all its power and might.

Whoever they put in charge of The Project is going to have a hell of a task: to build AGI, and to build it fast; to put the American economy on wartime footing to make hundreds of millions of GPUs; to lock it all down, weed out the spies, and fend off all-out attacks by the CCP; to somehow manage a hundred million AGIs furiously automating AI research, making a decade's leaps in a year, and soon producing AI systems *vastly* smarter than the smartest humans; to somehow keep things together enough that this doesn't go off the rails and produce rogue superintelligence that tries to seize control from its human overseers; to use those superintelligences to develop whatever new technologies will be necessary to stabilize the situation and stay ahead of adversaries, rapidly remaking US forces to integrate those; all while navigating what will likely be the tensest international situation ever seen. They better be good, I'll say that.

For those of us who get the call to come along for the ride, it'll be ... stressful. But it will be our duty to serve the free

Figure 38: Oppenheimer and Groves.

world—and all of humanity. If we make it through and get to look back on those years, it will be the most important thing we ever did. And while whatever secure facility they find probably won't have the pleasantries of today's ridiculously-overcomped-AI-researcher-lifestyle, it won't be so bad. SF already feels like a peculiar AI-researcher-college-town; probably this won't be so different. It'll be the same weirdly-small circle sweating the scaling curves during the day and hanging out over the weekend, kibitzing over AGI and the lab-politics-of-the-day.

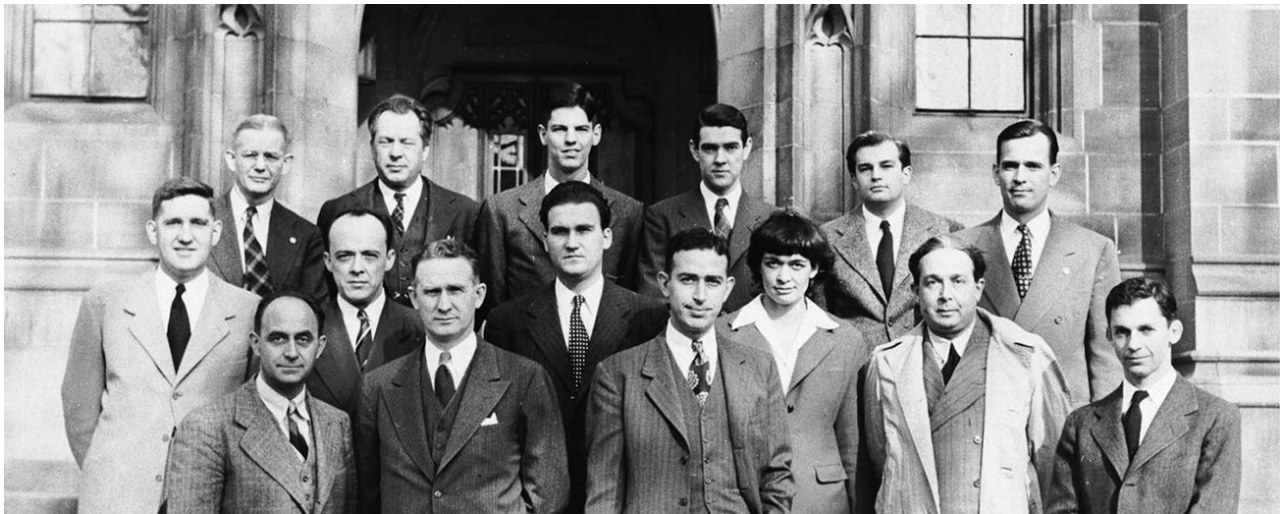Except, well—the stakes will be all too real.

See you in the desert, friends.

Figure 39: Reunion of atomic scientists on the fourth anniversary of the first controlled nuclear fission reaction at UChicago.

V. Parting Thoughts

What if we're right?

> I remember the spring of 1941 to this day. I realized then that a nuclear bomb was not only possible — it was inevitable. Sooner or later these ideas could not be peculiar to us. Everybody would think about them before long, and some country would put them into action. . . .
>
> And there was nobody to talk to about it, I had many sleepless nights. But I did realize how very very serious it could be. And I had then to start taking sleeping pills. It was the only remedy, I've never stopped since then. It's 28 years, and I don't think I've missed a single night in all those 28 years.
>
> <div align="right">

JAMES CHADWICH
(Physics Nobel Laureate and author of the 1941 British government report on the inevitability of an atomic bomb, which finally spurred the Manhattan Project into action)
</div>

BEFORE THE DECADE IS OUT, we will have built superintelligence. That is what most of this series has been about. For most people I talk to in SF, that's where the screen goes black. But the decade after—the 2030s—will be at least as eventful. By the end of it, the world will have been utterly, unrecognizably transformed. A new world order will have been forged. But alas—that's a story for another time.

We must come to a close, for now. Let me make a few final remarks.

AGI realism

This is all much to contemplate—and many cannot. "Deep learning is hitting a wall!" they proclaim, every year. It's just another tech boom, the pundits say confidently. But even among those at the SF-epicenter, the discourse has become polarized between two fundamentally unserious rallying cries.

On the one end there are the doomers. They have been obsessing over AGI for many years; I give them a lot of credit for their prescience. But their thinking has become ossified, untethered from the empirical realities of deep learning, their proposals naive and unworkable, and they fail to engage with the very real authoritarian threat. Rabid claims of 99% odds of doom, calls to indefinitely pause AI—they are clearly not the way.

On the other end are the e/accs. Narrowly, they have some good points: AI progress must continue. But beneath their shallow Twitter shitposting, they are a sham; dilettantes who just want to build their wrapper startups rather than stare AGI in the face. They claim to be ardent defenders of American freedom, but can't resist the siren song of unsavory dictators' cash. In truth, they are real stagnationists. In their attempt to deny the risks, they deny AGI; essentially, all we'll get is cool chatbots, which surely aren't dangerous. (That's some underwhelming accelerationism in my book.)

But as I see it, the smartest people in the space have converged on a different perspective, a third way, one I will dub **AGI Realism.** The core tenets are simple:

1. *Superintelligence is a matter of national security*. We are rapidly building machines smarter than the smartest humans. This is not another cool Silicon Valley boom; this isn't some random community of coders writing an innocent open source software package; this isn't fun and games. Superintelligence is going to be *wild*; it will be the most powerful weapon mankind has ever built. And for any of us involved, it'll be the most important thing we ever do.

2. ***America must lead.*** The torch of liberty will not survive Xi getting AGI first. (And, realistically, American leadership is the only path to safe AGI, too.) That means we can't simply "pause"; it means we need to rapidly scale up US power production to build the AGI clusters in the US. But it also means amateur startup security delivering the nuclear secrets to the CCP won't cut it anymore, and it means the core AGI infrastructure must be controlled by America, not some dictator in the Middle East. American AI labs must put the national interest first.

3. ***We need to not screw it up.*** Recognizing the power of superintelligence also means recognizing its peril. There are very real safety risks; very real risks this all goes awry—whether it be because mankind uses the destructive power brought forth for our mutual annihilation, or because, yes, the alien species we're summoning is one we cannot yet fully control. These are manageable—but improvising won't cut it. Navigating these perils will require good people bringing a level of seriousness to the table that has not yet been offered.

As the acceleration intensifies, I only expect the discourse to get more shrill. But my greatest hope is that there will be those who feel the weight of what is coming, and take it as a solemn call to duty.

What if we're right?

At this point, you may think that I and all the other SF-folk are totally crazy. But consider, just for a moment: *what if they're right?* These are the people who invented and built this technology; they think AGI will be developed this decade; and, though there's a fairly wide spectrum, many of them take very seriously the possibility that the road to superintelligence will play out as I've described in this series.

Almost certainly I've gotten important parts of the story wrong; if reality turns out to be anywhere near this crazy, the error bars will be very large. Moreover, as I said at the outset, I think there's a wide range of possibilities. But I think it is impor-

tant to be concrete. And in this series I've laid out what I currently believe is the single most likely scenario for the rest of the decade—the rest of *this* decade.

Because—it's starting to feel real, *very* real. A few years ago, at least for me, I took these ideas seriously—but they were abstract, quarantined in models and probability estimates. Now it feels extremely visceral. I can *see* it. I can *see* how AGI will be built. It's no longer about estimates of human brain size and hypotheticals and theoretical extrapolations and all that—I can basically tell you the cluster AGI will be trained on and when it will be built, the rough combination of algorithms we'll use, the unsolved problems and the path to solving them, the list of people that will matter. I can *see* it. It is extremely visceral. Sure, going all-in leveraged long Nvidia in early 2023 has been great and all, but the burdens of history are heavy. I would not choose this.

But the scariest realization is that *there is no crack team coming to handle this*. As a kid you have this glorified view of the world, that when things get real there are the heroic scientists, the uber-competent military men, the calm leaders who are on it, who will save the day. It is not so. The world is incredibly small; when the facade comes off, it's usually just a few folks behind the scenes who are the live players, who are desperately trying to keep things from falling apart.

Right now, there's perhaps a few hundred people in the world who realize what's about to hit us, who understand just how crazy things are about to get, who have situational awareness. I probably either personally know or am one degree of separation from everyone who could plausibly run The Project. The few folks behind the scenes who are desperately trying to keep things from falling apart are you and your buddies and their buddies. That's it. That's all there is.

Someday it will be out of our hands. But right now, at least for the next few years of midgame, the fate of the world rests on these people.

Will the free world prevail?

Will we tame superintelligence, or will it tame us?

Will humanity skirt self-destruction once more?

The stakes are no less.

These are great and honorable people. But they are just people. Soon, the AIs will be running the world, but we're in for one last rodeo. May their final stewardship bring honor to mankind.

Appendix

Some additional details on compute back-of-the-envelope calculca-tions

Training compute

Year.   The OpenAI GPT-4 tech report stated that GPT-4 fin-ished training in August 2022. Thereafter we play forward the rough 0.5 OOMs/year trend.

H100s-equivalent.   Semianalysis, JP Morgan, and others esti-mate GPT-4 was trained on 25k A100s, and H100s are 2-3x the performance of A100s.

Cost.   Often people cite numbers like "$100M for GPT-4 train-ing," using just the rental cost of the GPUs (i.e., something like "how much would it cost to rent this size cluster for 3 months of training"). But that's a mistake. What matters is something more like ~the actual cost to build the cluster. If you want one of the largest clusters in the world, you can't just rent it for 3 months! And moreover, you need the compute for more than just the flagship training run: there will be lots of derisking experiments, failed runs, other models, etc.

To approximate the GPT-4 cluster cost:

- The public estimates suggest the GPT-4 cluster being around 25k A100s.

- Assuming $1/A100-hour for 2-3 years gives roughly a $500M cost.

- Alternatively, you can estimate it as $25k cost per H100, 10k H100s-equivalent, and Nvidia GPUs being around half the cost of a cluster (the rest being power, the physical datacenter, cooling, networking, maintenance personnel, etc.). (For example, this total-cost-of-ownership analysis estimates that around 40% of a large cluster cost is the H100 GPUs itself, and another 13% goes to Nvidia for Infiniband networking. That said, excluding cost of capital in that calculation would mean the GPUs are about 50% of the cost, and with networking Nvidia gets a bit over 60% of the cost of the cluster.)

FLOP/$ is improving somewhat for each Nvidia generation, but not a ton. E.g., the H100 -> B100 is likely something like 1.5x FLOP/$: B100s are effectively two H100s stapled together, but retailing for <2x the cost. But the B100 was somewhat of an exception. They are surprisingly cheap, likely because Nvidia wants to crush competition. By contrast, A100s -> H100s weren't much of a FLOP/$ improvement (2x better chip without fp8, roughly 2x the cost), maybe 1.5x if we count fp8 improvements—and that was for a two-year generation.

While I think there are some tailwinds to further FLOP/$ improvements from margin compression, GPUs might also get more expensive as they become massively constrained. Gains from AI chip specialization will continue, but it's not clear to me there will still be game-changing technical improvements to FLOP/$ coming, given chips are already pretty specialized for AI (e.g. specialized for Transformers, and already at fp8/fp4 precision), Moore's Law is glacial these days, and other bottleneck components like memory and interconnect are improving more slowly. If you look at Epoch's data, there seems to be less than a 10x in FLOP/$ over the past decade for top ML GPUs, and for the aforementioned reasons if anything I'd expect this to slow down.

Something like a 35%/year improvement in FLOP/$ would give us the 1T cost for the +4 OOM cluster. Maybe FLOP/$ improves faster, but also datacenter capex is going to get more

expensive—simply because you'll need to actually build new power, doing a lot of capex up front, rather than just renting existing depreciated power plants.

These are just very rough numbers anyway. It would be very much within error bars if e.g. the 1T cluster can be done more efficiently and actually yields more like +4.5 OOMs on compute.

Cost.   An H100 is 700W, but there's a bunch of datacenter power you need (cooling, networking, storage); Semianalysis estimates ~1,400W per H100.

There are some gains to be had on FLOP/Watt, though once we've exhausted the gains from AI chip specialization (see previous footnote), e.g. gone down to the lowest possible precision, these seem somewhat limited (mostly just chip process improvements, which are slow). That said, as power becomes more of a constraint (and thus a larger fraction of costs), chip designs might specialize to more power-efficiency at the costs of FLOPs. Still, there is still the power demand for cooling, networking, storage, and so on (which in the H100 numbers above was already roughly half the power demand).

For these back of the envelope numbers, let's work with 1kW per H100-equivalent here; again these are just rough back-of-the-envelope calculations. (And if there were an unexpected FLOP/Watt breakthrough, I'd expect the same power expenditure, just bigger OOM compute gains.)

Power reference classes.   A 10GW cluster run continuously for a year is 87.6 TWh. By comparison, Oregon consumes about 27 TWh of electricity annually, Washington state consumes about 92 TWh annually.

A 100GW cluster run continuously for a year is 876 TWh, while total annual US electricity production is about 4,250 TWh.

Overall compute

H100-equivalents this year.   I estimate Nvidia is going to ship on the order of 5M datacenter GPUs in 2024. A minority of those are B100s, which we'll count as 2x+ H100s. Then there's the other AI chips: TPUs, Trainium, Meta's custom silicon, AMD GPUs, etc.

TSMC Capacity.   TSMC has capacity for over 150k 5nm wafers per month, is ramping to 100k 3nm wafers per month, and likely another 150k or so 7nm wafers per month; let's call it 400k wafers per month in total.

Let's say roughly 35 H100s per wafer (H100s are made on 5nm). At 5-10 million H100-equivalents in 2024, that's 150k-300k wafers per year for annual AI chip production in 2024.

Depending on where in that range and whether we want to count 7nm production, that's about 3-10% of annual leading-edge wafer production.

Click here to return back to the "overall compute" table.