Dwarkesh Podcast #74 - Leopold Aschenbrenner - China/US Super Intelligence Race, 2027 AGI, & The Return of History

Published - June 4, 2024

Transcribed by - thepodtranscripts.com

Dwarkesh Patel
Today I'm chatting with my friend Leopold Aschenbrenner. He grew up in Germany and graduated as valedictorian of Columbia when he was 19. After that, he had a very interesting gap year which we'll talk about. Then, he was on the OpenAI superalignment team, may it rest in peace.

Now, with some anchor investments — from Patrick and John Collison, Daniel Gross, and Nat Friedman — he is launching an investment firm.

Leopold, you're off to a slow start but life is long. I wouldn't worry about it too much. You'll make up for it in due time. Thanks for coming on the podcast.

Leopold Aschenbrenner
Thank you. I first discovered your podcast when your best episode had a couple of hundred views. It's been amazing to follow your trajectory. It's a delight to be on.

Dwarkesh Patel
In the Sholto and Trenton episode, I mentioned that a lot of the things I've learned about AI I've learned from talking with them. The third, and probably most significant, part of this triumvirate has been you. We'll get all the stuff on the record now.

Here's the first thing I want to get on the record. Tell me about the trillion-dollar cluster.

I should mention this for the context of the podcast. Today you're releasing a series called Situational Awareness. We're going to get into it. First question about that is, tell me about the trillion-dollar cluster.

Leopold Aschenbrenner
Unlike most things that have recently come out of Silicon Valley, AI is an industrial process. The next model doesn't just require some code. It's building a giant new cluster. It's building giant new power plants. Pretty soon, it's going to involve building giant new fabs.

Since ChatGPT, this extraordinary techno-capital acceleration has been set into motion. Exactly a year ago today, Nvidia had their first blockbuster earnings call. It went up 25% after hours and everyone was like, "oh my God, AI is a thing." Within a year, Nvidia data center revenue has gone from a few billion a quarter to $25 billion a quarter and continues to go up. Big Tech capex is skyrocketing.

It's funny. There's this crazy scramble going on, but in some sense it's just the continuation of straight lines on a graph. There's this long-run trend of almost a decade of training compute for the largest AI systems growing by about half an order of magnitude, 0.5 OOMs a year.

Just play that forward. GPT-4 was reported to have finished pre-training in 2022. On SemiAnalysis, it was rumored to have a cluster size of about 25,000 A100s. That's roughly a $500 million cluster. Very roughly, it's 10 megawatts.

Just play that forward half a year. By 2024, that's a cluster that's 100 MW and 100,000 H100 equivalents with costs in the billions.

Play it forward two more years. By 2026, that's a gigawatt, the size of a large nuclear reactor. That's like the power of the Hoover Dam. That costs tens of billions of dollars and requires a million H100 equivalents.

By 2028, that's a cluster that's ten GW. That's more power than most US states. That's 10 million H100 equivalents, costing hundreds of billions of dollars.

By 2030, you get the trillion-dollar cluster using 100 gigawatts, over 20% of US electricity production. That's 100 million H100 equivalents.

That's just the training cluster. There are more inference GPUs as well. Once there are products, most of them will be inference GPUs. US power production has barely grown for decades. Now we're really in for a ride.

Dwarkesh Patel
When I had Zuck on the podcast, he was claiming not a plateau per se, but that AI progress would be bottlenecked by this constraint on energy. Specifically, he was like, "oh, gigawatt data centers, are we going to build another Three Gorges Dam or something?"

According to public reports, there are companies planning things on the scale of a 1 GW data center. With a 10 GW data center, who's going to be able to build that? A 100 GW center is like a state project. Are you going to pump that into one physical data center? How is it going to be possible? What is Zuck missing?

Leopold Aschenbrenner
Six months ago, 10 GW was the talk of the town. Now, people have moved on. 10 GW is happening. There's The Information report on OpenAI and Microsoft planning a $100 billion cluster.

Dwarkesh Patel
Is that 1 GW? Or is that 10 GW?

Leopold Aschenbrenner
I don't know but if you try to map out how expensive the 10 GW cluster would be, that's a couple of hundred billion. It's sort of on that scale and they're planning it. It's not just my

crazy take. AMD forecasted a $400 billion AI accelerator market by 2027. AI accelerators are only part of the expenditures.

We're very much on track for a $1 trillion of total AI investment by 2027. The $1 trillion cluster will take a bit more acceleration. We saw how much ChatGPT unleashed. Every generation, the models are going to be crazy and shift the Overton window.

Then the revenue comes in. These are forward-looking investments. The question is, do they pay off? Let's estimate the GPT-4 cluster at around $500 million. There's a common mistake people make, saying it was $100 million for GPT-4. That's just the rental price. If you're building the biggest cluster, you have to build and pay for the whole cluster. You can't just rent it for three months.

Dwarkesh Patel
Can't you?

Leopold Aschenbrenner
Once you're trying to get into the hundreds of billions, you have to get to like $100 billion a year in revenue. This is where it gets really interesting for the big tech companies because their revenues are on the order of hundreds of billions.

$10 billion is fine. It'll pay off the 2024 size training cluster. It'll really be gangbusters with Big Tech when it costs $100 billion a year. The question is how feasible is $100 billion a year from AI revenue? It's a lot more than right now. If you believe in the trajectory of AI systems as I do, it's not that crazy.

There are like 300 million Microsoft Office subscribers. They have Copilot now. I don't know what they're selling it for. Suppose you sold some AI add-on for $100/month to a third of Microsoft Office subscribers. That'd be $100 billion right there. $100/month is a lot.

Dwarkesh Patel
That's a lot for a third of Office subscribers.

Leopold Aschenbrenner
For the average knowledge worker, it's a few hours of productivity a month. You have to be expecting pretty lame AI progress to not hit a few hours of productivity a month.

Dwarkesh Patel
Sure, let's assume all this. What happens in the next few years? What can the AI trained on the 1 GW data center do? What about the one on the 10 GW data center? Just map out the next few years of AI progress for me.

Leopold Aschenbrenner
The 10 GW range is my best guess for when you get true AGI. Compute is actually overrated. We'll talk about that.

By 2025-2026, we're going to get models that are basically smarter than most college graduates. A lot of the economic usefulness depends on unhobbling. The models are smart but limited. There are chatbots and then there are things like being able to use a computer and doing agentic long-horizon tasks.

By 2027-2028, it'll get as smart as the smartest experts. The unhobbling trajectory points to it becoming much more like an agent than a chatbot. It'll almost be like a drop-in remote worker.

This is the question around the economic returns. Intermediate AI systems could be really useful, but it takes a lot of schlep to integrate them. There's a lot you could do with GPT-4 or GPT-4.5 in a business use case, but you really have to change your workflows to make them useful. It's a very Tyler Cowen-esque take. It just takes a long time to diffuse. We're in SF and so we miss that.

But in some sense, the way these systems want to be integrated is where you get this kind of sonic boom. Intermediate systems could have done it, but it would have taken schlep. Before you do the schlep to integrate them, you'll get much more powerful systems that are unhobbled.

They're agents, drop-in remote workers. You're interacting with them like coworkers. You can do Zoom calls and Slack with them. You can ask them to do a project and they go off and write a first draft, get feedback, run tests on their code, and come back. Then you can tell them more things. That'll be much easier to integrate.

You might need a bit of overkill to make the transition easy and harvest the gains.

Dwarkesh Patel
What do you mean by overkill? Overkill on model capabilities?

Leopold Aschenbrenner
Yeah, the intermediate models could do it but it would take a lot of schlep. The drop-in remote worker AGI can automate cognitive tasks. The intermediate models would have made the software engineer more productive. But will the software engineer adopt it?

With the 2027 model, you just don't need the software engineer. You can interact with it like a software engineer, and it'll do the work of a software engineer.

Dwarkesh Patel
The last episode I did was with John Schulman.

I was asking about this. We have these models that have come out in the last year and none seem to have significantly surpassed GPT-4, certainly not in an agentic way where they interact with you as a coworker. They'll brag about a few extra points on MMLU. Even with GPT-4o, it's cool they can talk like Scarlett Johansson (I guess not anymore) but it's not like a coworker.

It makes sense why they'd be good at answering questions. They have data on how to complete Wikipedia text. Where is the equivalent training data to understand a Zoom call? Referring back to your point about a Slack conversation, how can it use context to figure out the cohesive project you're working on? Where is that training data coming from?

Leopold Aschenbrenner
A key question for AI progress in the next few years is how hard it is to unlock the test time compute overhang. Right now, GPT-4 can do a few hundred tokens with chain-of-thought. That's already a huge improvement. Before, answering a math question was just shotgun. If you tried to answer a math question by saying the first thing that comes to mind, you wouldn't be very good.

GPT-4 thinks for a few hundred tokens. If I think at 100 tokens a minute, that's like what GPT-4 does. It's equivalent to me thinking for three minutes. Suppose GPT-4 could think for millions of tokens. That's +4 OOMs on test time compute on one problem. It can't do it now. It gets stuck. It writes some code. It can do a little bit of iterative debugging, but eventually gets stuck and can't correct its errors.

There's a big overhang. In other areas of ML, there's a great paper on AlphaGo, where you can trade off train time and test time compute. If you can use 4 OOMs more test time compute, that's almost like a 3.5x OOM bigger model.

Again, if it's 100 tokens a minute, a few million tokens is a few months of working time. There's a lot more you can do in a few months of working time than just getting an answer right now. The question is how hard is it to unlock that?

In the short timelines AI world, it's not that hard. The reason it might not be that hard is that there are only a few extra tokens to learn. You need to learn things like error correction tokens where you're like "ah, I made a mistake, let me think about that again." You need to learn planning tokens where it's like "I'm going to start by making a plan. Here's my plan of attack. I'm going to write a draft and now I'm going to critique my draft and think about it." These aren't things that models can do now, but the question is how hard it is.

There are two paths to agents. When Sholto was on your podcast, he talked about scaling leading to more nines of reliability. That's one path. The other path is the unhobbling path. It needs to learn this System 2 process. If it can learn that, it can use millions of tokens and think coherently.

Here's an analogy. When you drive, you're on autopilot most of the time. Sometimes you hit a weird construction zone or intersection. Sometimes my girlfriend is in the passenger seat and I'm like "ah, be quiet for a moment, I need to figure out what's going on."

You go from autopilot to System 2 and you're thinking about how to do it. Scaling improves that System 1 autopilot. The brute force way to get to agents is improving that system. If you can get System 2 working, you can quickly jump to something more agentified and test time compute overhang is unlocked.

Dwarkesh Patel
What's the reason to think this is an easy win? Is there some loss function that easily enables System 2 thinking? There aren't many animals with System 2 thinking. It took a long time for evolution to give us System 2 thinking.

Pre-training has trillions of tokens of Internet text, I get that. You match that and get all of these free training capabilities. What's the reason to think this is an easy unhobbling?

Leopold Aschenbrenner
First of all, pre-training is magical. It gave us a huge advantage for models of general intelligence because you can predict the next token. But there's a common misconception. Predicting the next token lets the model learn incredibly rich representations.

Representation learning properties are the magic of deep learning. Rather than just learning statistical artifacts, the models learn models of the world. That's why they can generalize, because it learned the right representations.

When you train a model, you have this raw bundle of capabilities that's useful. The unhobbling from GPT-2 to GPT-4 took this raw mass and RLHF'd it into a good chatbot. That was a huge win.

In the original InstructGPT paper, comparing RLHF vs. non-RLHF models it's like a 100x model size win on human preference rating. It started to be able to do simple chain-of-thought and so on. But you still have this advantage of all these raw capabilities, and there's still a huge amount you're not doing with them.

This pre-training advantage is also the difference to robotics. People used to say it was a hardware problem. The hardware is getting solved, but you don't have this huge advantage of bootstrapping with pre-training. You don't have all this unsupervised learning you can do. You have to start right away with RL self-play.

The question is why RL and unhobbling might work. Bootstrapping is an advantage. Your Twitter bio is being pre-trained. You're not being pre-trained anymore. You were pre-trained in grade school and high school. At some point, you transition to being able to learn by yourself. You weren't able to do it in elementary school. High school is probably where it started and by college, if you're smart, you can teach yourself. Models are just starting to enter that regime.

It's a little bit more scaling and then you figure out what goes on top. It won't be trivial. A lot of deep learning seems obvious in retrospect. There's some obvious cluster of ideas. There are some ideas that seem a little dumb but work. There are a lot of details you have to get right. We're not going to get this next month. It'll take a while to figure out.

Dwarkesh Patel
A while for you is like half a year.

Leopold Aschenbrenner
I don't know, between six months and three years. But it's possible. It's also very related to the issue of the data wall. Here's one intuition on learning by yourself. Pre-training is kind of like the teacher lecturing to you and the words are flying by. You're just getting a little bit from it.

That's not what you do when you learn by yourself. When you learn by yourself, say you're reading a dense math textbook, you're not just skimming through it once. Some wordcels just skim through and reread and reread the math textbook and they memorize.

What you do is you read a page, think about it, have some internal monologue going on, and have a conversation with a study buddy. You try a practice problem and fail a bunch of times. At some point it clicks, and you're like, "this made sense." Then you read a few more pages.

We've kind of bootstrapped our way to just starting to be able to do that now with models. The question is, can you use all this sort of self-play, synthetic data, RL to make that thing work. Right now, there's in-context learning, which is super sample efficient. In the Gemini paper, it just learns a language in-context. Pre-training, on the other hand, is not at all sample efficient.

What humans do is a kind of in-context learning. You read a book, think about it, until eventually it clicks. Then you somehow distill that back into the weights. In some sense, that's what RL is trying to do. RL is super finicky, but when it works it's kind of magical.

It's the best possible data for the model. It's when you try a practice problem, fail, and at some point figure it out in a way that makes sense to you. That's the best possible data for you because it's the way you would have solved the problem, rather than just reading how somebody else solved the problem, which doesn't initially click.

Dwarkesh Patel
By the way, if that take sounds familiar it's because it was part of the question I asked John Schulman. It goes to illustrate the thing I said in the intro. A bunch of the things I've learned about AI comes from these dinners we do before the interviews with me, you, Sholto, and a couple of others. We're like, "what should I ask John Schulman, what I should ask Dario." Suppose this is the way things go and we get these unhobblings —

Leopold Aschenbrenner
And the scaling. You have this baseline of this enormous force of scaling. GPT-2 was amazing. It could string together plausible sentences, but it could barely do anything. It was kind of like a preschooler. GPT-4, on the other hand, could write code and do hard math, like a smart high schooler. This big jump in capability is explored in the essay series. I count the orders of magnitude of compute and scale-up of algorithmic progress.

Scaling alone by 2027-2028 is going to do another preschool to high school jump on top of GPT-4. At a per token level, the models will be incredibly smart. They'll gain more reliability, and with the addition of unhobblings, they'll look less like chatbots and more like agents or drop-in remote workers. That's when things really get going.

Dwarkesh Patel
I want to ask more questions about this but let's zoom out. Suppose you're right about this. This is because of the 2027 cluster which is at 10 GW?

Leopold Aschenbrenner
2028 is 10 GW. Maybe it'll be pulled forward.

Dwarkesh Patel
Something like a 5.5 level by 2027, whatever that's called. What does the world look like at that point? You have these remote workers who can replace people. What is the reaction to that in terms of the economy, politics, and geopolitics?

Leopold Aschenbrenner
2023 was a really interesting year to experience as somebody who was really following the AI stuff.

Dwarkesh Patel
What were you doing in 2023?

Leopold Aschenbrenner
OpenAI. When you were at OpenAI in 2023, it was a weird thing. You almost didn't want to talk about AI or AGI. It was kind of a dirty word. Then in 2023, people saw ChatGPT for the first time, they saw GPT-4, and it just exploded.

It triggered huge capital expenditures from all these firms and an explosion in revenue from Nvidia and so on. Things have been quiet since then, but the next thing has been in the oven. I expect every generation these g-forces to intensify. People will see the models. They won’t have counted the OOMs so they're going to be surprised. It'll be kind of crazy.

Revenue is going to accelerate. Suppose you do hit $10 billion by the end of this year. Suppose it just continues on the trajectory of revenue doubling every six months. It's not actually that far from $100 billion, maybe by 2026. At some point, what happened to Nvidia is going to happen to Big Tech. It's going to explode. A lot more people are going to feel it.

2023 was the moment for me where AGI went from being this theoretical, abstract thing. I see it, I feel it, and I see the path. I see where it's going. I can see the cluster it's trained on, the rough combination of algorithms, the people, how it's happening. Most of the world is not there yet. Most of the people who feel it are right here. A lot more of the world is going to start feeling it. That's going to start being intense.

Dwarkesh Patel
Right now, who feels it? You can go on Twitter and there are these GPT-wrapper companies, like, "whoa, GPT-4 is going to change our business."

Leopold Aschenbrenner
I'm so bearish on the wrapper companies because they're betting on stagnation. They're betting that you have these intermediate models and it takes so much schlep to integrate them. I'm really bearish because we're just going to sonic boom you. We're going to get the unhobblings. We're going to get the drop-in remote worker. Your stuff is not going to matter.

Dwarkesh Patel
So that's done. SF, this crowd, is paying attention now. Who is going to be paying attention in 2026 and 2027? Presumably, these are years in which hundreds of billions of capex is being spent on AI.

Leopold Aschenbrenner
The national security state is going to start paying a lot of attention. I hope we get to talk about that.

Dwarkesh Patel
Let’s talk about it now. What happens? What is the immediate political reaction? Looking internationally, I don't know if Xi Jinping sees the GPT-4 news and goes, "Oh, my God. Look at the MMLU score on that. What are we doing about this, comrade?"

So what happens when he sees a remote worker replacement and it has $100 billion in revenue? There’s a lot of businesses that have $100 billion in revenue, and people aren't staying up all night talking about it.

Leopold Aschenbrenner
The question is, when does the CCP and when does the American national security establishment realize that superintelligence is going to be absolutely decisive for national power? This is where the intelligence explosion stuff comes in, which we should talk about later.

You have AGI. You have this drop-in remote worker that can replace you or me, at least for remote jobs. Fairly quickly, you turn the crank one or two more times and you get a thing that's smarter than humans.

Even more than just turning the crank a few more times, one of the first jobs to be automated is going to be that of an AI researcher or engineer. If you can automate AI research, things can start going very fast.

Right now, there's already at this trend of 0.5 OOMs a year of algorithmic progress. At some point, you're going to have GPU fleets in the tens of millions for inference or more. You’re going to be able to run 100 million human equivalents of these automated AI researchers.

If you can do that, you can maybe do a decade's worth of ML research progress in a year. You get some sort of 10x speed up. You can make the jump to AI that is vastly smarter than humans within a year, a couple of years.

That broadens from there. You have this initial acceleration of AI research. You apply R&D to a bunch of other fields of technology. At this point, you have a billion super intelligent researchers, engineers, technicians, everything. They’re superbly competent at all things.

They're going to figure out robotics. We talked about that being a software problem. Well, you have a billion super smart — smarter than the smartest human researchers — AI researchers in your cluster. At some point during the intelligence explosion, they're going to be able to figure out robotics. Again, that’ll expand.

If you play this picture forward, it is fairly unlike any other technology. A couple years of lead could be utterly decisive in say, military competition. If you look at the first Gulf War, Western coalition forces had a 100:1 kill ratio. They had better sensors on their tanks. They had better precision missiles, GPS, and stealth. They had maybe 20-30 years of technological lead. They just completely crushed them.

Superintelligence applied to broad fields of R&D — and the industrial explosion that comes from it, robots making a lot of material — could compress a century’s worth of technological progress into less than a decade. That means that a couple years could mean a Gulf War 1-style advantage in military affairs. That’s including a decisive advantage that even preempts nukes.

How do you find nuclear stealth submarines? Right now, you have sensors and software to detect where they are. You can do that. You can find them. You have millions or billions of mosquito-sized drones, and they take out the nuclear submarines. They take out the mobile launchers. They take out the other nukes.

It’s potentially enormously destabilizing and enormously important for national power. At some point people are going to realize that. Not yet, but they will. When they do, it won’t just be the AI researchers in charge.

The CCP is going to have an all-out effort to infiltrate American AI labs. It’ll involve billions of dollars, thousands of people, and the full force of the Ministry of State Security. The CCP is going to try to outbuild us.

They added as much power in the last decade as an entire US electric grid. So the 100 GW cluster, at least the 100 GW part of it, is going to be a lot easier for them to get. By this point, it's going to be an extremely intense international competition.

Dwarkesh Patel
One thing I'm uncertain about in this picture is if it’s like what you say, where it's more of an explosion. You’ve developed an AGI. You make it into an AI researcher. For a while, you're only using this ability to make hundreds of millions of other AI researchers. The thing that comes out of this really frenetic process is a superintelligence. Then that goes out in the world and is developing robotics and helping you take over other countries and whatever.

Leopold Aschenbrenner
It's a little bit more gradual. It's an explosion that starts narrowly. It can do cognitive jobs. The highest ROI use for cognitive jobs is to make the AI better and solve robotics. As you solve robotics, now you can do R&D in biology and other technology.

Initially, you start with the factory workers. They're wearing the glasses and AirPods, and the AI is instructing them because you can make any worker into a skilled technician. Then you have the robots come in. So this process expands.

Dwarkesh Patel
Meta's Ray-Bans are a complement to Llama.

Leopold Aschenbrenner
With the fabs in the US, their constraint is skilled workers. Even if you don't have robots, you have the cognitive superintelligence and can kind of make them all into skilled workers immediately. That's a very brief period. Robots will come soon.

Dwarkesh Patel
Suppose this is actually how the tech progresses in the United States, maybe because these companies are already generating hundreds of billions of dollars of AI revenue

Leopold Aschenbrenner
At this point, companies are borrowing hundreds of billions or more in the corporate debt markets.

Dwarkesh Patel
Why is a CCP bureaucrat, some 60-year-old guy, looking at this and going, "Oh, Copilot has gotten better now." And now —

Leopold Aschenbrenner
This is much more than Copilot has gotten better now.

Dwarkesh Patel
It'd require shifting the production of an entire country, dislocating energy that is otherwise being used for consumer goods or something, and feeding all that into the data centers. Part of this whole story is that you realize superintelligence is coming soon. You realize it and maybe I realize it. I'm not sure how much I realize it.

Will the national security apparatus in the United States and the CCP realize it?

Leopold Aschenbrenner
This is a really key question. We have a few more years of mid-game. We have a few more 2023s. That just starts updating more and more people. The trend lines will become clear.

You will see some amount of the COVID dynamic. COVID in February of 2020 honestly feels a lot like today. It feels like this utterly crazy thing is coming. You see the exponential and yet most of the world just doesn't realize it. The mayor of New York is like, "go out to the shows," and "this is just Asian racism." At some point, people saw it and then crazy, radical reactions came.

Dwarkesh Patel
By the way, what were you doing during COVID? Was it your freshman or sophomore year?

Leopold Aschenbrenner
Junior.

Dwarkesh Patel
Still, you were like a 17-year-old junior or something right? Did you short the market or something? Did you sell at the right time?

Leopold Aschenbrenner
Yeah.

Dwarkesh Patel
So there will be a March 2020 moment.

You can make the analogy you make in the series that this will cause a reaction like, “we have to do the Manhattan Project again for America here.” I wonder what the politics of this will be like. The difference here is that it’s not just like, “we need the bomb to beat the Nazis.”

We'll be building this thing that makes all our energy prices go up a bunch and it's automating a lot of our jobs. The climate change stuff people are going to be like, "oh, my God, it's making climate change worse and it's helping Big Tech."

Politically, this doesn't seem like a dynamic where the national security apparatus or the president is like, "we have to step on the gas here and make sure America wins."

Leopold Aschenbrenner
Again, a lot of this really depends on how much people are feeling it and how much people are seeing it. Our generation is so used to peace, American hegemony and nothing matters. The historical norm is very much one of extremely intense and extraordinary things happening in the world with intense international competition.

There's a 20-year very unique period. In World War II, something like 50% of GDP went to war production. The US borrowed over 60% of GDP. With Germany and Japan I think it was over 100%. In World War I, the UK, France, and Germany all borrowed over 100% of GDP.

Much more was on the line. People talk about World War I being so destructive with 20 million Soviet soldiers dying and 20% of Poland. That happened all the time. During the Seven Years' War something like 20-30% of Prussia died. In the Thirty Years' War, up to 50% of a large swath of Germany died.

Will people see that the stakes here are really high and that history is actually back? The American national security state thinks very seriously about stuff like this. They think very seriously about competition with China. China very much thinks of itself on this historical mission of the rejuvenation of the Chinese nation. They think a lot about national power. They think a lot about the world order.

There's a real question on timing. Do they start taking this seriously when the intelligence explosion is already happening quite late. Do they start taking this seriously two years earlier? That matters a lot for how things play out.

At some point they will and they will realize that this will be utterly decisive for not just some proxy war but for major questions. Can liberal democracy continue to thrive? Can the CCP continue existing? That will activate forces that we haven't seen in a long time.

Dwarkesh Patel
The great power conflict definitely seems compelling. All kinds of different things seem much more likely when you think from a historical perspective. You zoom out beyond the liberal democracy that we've had the pleasure to live in America for say the last 80 years. That includes things like dictatorships, war, famine, etc.

I was reading The Gulag Archipelago and one of the chapters begins with Solzhenitsyn saying how if you had told a Russian citizen under the tsars that because of all these new technologies — we wouldn't see some Great Russian revival with Russia becoming a great power and the citizens made wealthy — you would see tens of millions of Soviet citizens tortured by millions of beasts in the worst possible ways. If you'd told them that that would be the result of the 20th century, they wouldn't have believed you. They'd have called you a slanderer.

Leopold Aschenbrenner
The possibilities for dictatorship with superintelligence are even crazier as well. Imagine you have a perfectly loyal military and security force. No more rebellions. No more popular uprisings. You have perfect lie detection. You have surveillance of everybody. You can perfectly figure out who's the dissenter and weed them out. No Gorbachev who had some doubts about the system would have ever risen to power. No military coup would have ever happened.

There's a real way in which part of why things have worked out is that ideas can evolve. There's some sense in which time heals a lot of wounds and solves a lot of debates. Throughout time, a lot of people had really strong convictions, but a lot of those have been overturned over time because there's been continued pluralism and evolution.

Imagine applying a CCP-like approach to truth where truth is what the party says. When you supercharge that with superintelligence, that could just be locked in and enshrined for a long time. The possibilities are pretty terrifying.

To your point about history and living in America for the past eight years, this is one of the things I took away from growing up in Germany. A lot of this stuff feels more visceral. My mother grew up in the former East, my father in the former West. They met shortly after the Wall fell. The end of the Cold War was this extremely pivotal moment for me because it's the reason I exist.

I grew up in Berlin with the former Wall. My great-grandmother, who is still alive, is very important in my life. She was born in 1934 and grew up during the Nazi era. In World War II, she saw the firebombing of Dresden from this country cottage where they were as kids. Then she spent most of her life in the East German communist dictatorship.

She'd tell me about how Soviet tanks came when there was the popular uprising in 1954. Her husband was telling her to get home really quickly and get off the streets. She had a son who tried to ride a motorcycle across the Iron Curtain and then was put in a Stasi prison for a while. Finally, when she's almost 60, it was the first time she lived in a free country, and a wealthy country.

When I was a kid, the thing she always really didn't want me to do was get involved in politics. Joining a political party had very bad connotations for her. She raised me when I was young. So it doesn't feel that long ago. It feels very close.

Dwarkesh Patel
There's one thing I wonder about when we're talking today about the CCP. The people in China who will be doing their version of this project will be AI researchers who are somewhat Westernized. They'll either have gotten educated in the West or have colleagues in the West.

Are they going to sign up for the CCP project that's going to hand over control to Xi Jinping? What's your sense of that? Fundamentally, they're just people, right? Can't you convince them about the dangers of superintelligence?

Leopold Aschenbrenner
Will they be in charge though? In some sense, this is also the case in the US. This is like the rapidly depreciating influence of the lab employees. Right now, the AI lab employees have so much power. You saw this November event. It's so much power.

Both are going to get automated and they're going to lose all their power. It'll just be a few people in charge with their armies of automated AIs. It's also the politicians and the generals and the national security state. There are some of these classic scenes from the Oppenheimer movie. The scientists built it and then the bomb was shipped away and it was out of their hands.

It's good for lab employees to be aware of this. You have a lot of power now, but maybe not for that long. Use it wisely. I do think they would benefit from some more organs of representative democracy.

Dwarkesh Patel
What do you mean by that?

Leopold Aschenbrenner
In the OpenAI board events, employee power is exercised in a very direct democracy way. How some of that went about really highlighted the benefits of representative democracy and having some deliberative organs.

Dwarkesh Patel
Interesting. Let's go back to the $100 billion revenue question. The companies are trying to build clusters that are this big. Where are they building it? Say it's the amount of energy that would be required for a small or medium-sized US state. Does Colorado then get no power because it's happening in the United States? Is it happening somewhere else?

Leopold Aschenbrenner
This is the thing that I always find funny, when you talk about Colorado getting no power. The easy way to get the power would be to displace less economically useful stuff. Buy up the aluminum smelting plant that has a gigawatt. We're going to replace it with the data center because that's important. That's not actually happening because a lot of these power contracts are really locked in long-term. Also, people don't like things like this.

In practice what it requires, at least right now, is building new power. That might change. That's when things get really interesting, when it's like, “no, we're just dedicating all of the power to the AGI.”

So right now it's building new power. 10 GW is quite doable. It's like a few percent of US natural gas production. When you have the 10 GW training cluster, you have a lot more inference. 100 gigawatts is where it starts getting pretty wild. That's over 20% of US electricity production. It's pretty doable, especially if you're willing to go for natural gas.

It is incredibly important that these clusters are in the United States.

Dwarkesh Patel
Why does it matter that it's in the US?

Leopold Aschenbrenner
There are some people who are trying to build clusters elsewhere. There's a lot of free-flowing Middle Eastern money that's trying to build clusters elsewhere. This comes back to the national security question we talked about. Would you do the Manhattan Project in the UAE?

You can put the clusters in the US and you can put them in allied democracies. Once you put them in authoritarian dictatorships, you create this irreversible security risk. Once the cluster is there, it's much easier for them to exfiltrate the weights. They can literally steal the AGI, the superintelligence. It’s like they got a direct copy of the atomic bomb. It makes it much easier for them. They have weird ties to China. They can ship that to China. That's a huge risk.

Another thing is they can just seize the compute. The issue here is people right now are thinking of this as ChatGPT, Big Tech product clusters. The clusters being planned now, three to five years out, may well be the AGI, superintelligence clusters. When things get hot, they might just seize the compute.

Suppose we put 25% of the compute capacity in these Middle Eastern dictatorships. Say they seize that. Now it's a ratio of compute of 3:1. We still have more, but even with only 25% of compute there it starts getting pretty hairy. 3:1 is not that great of a ratio. You can do a lot with that amount of compute.

Say they don't actually do this. Even if they don't actually seize the compute, even if they actually don't steal the weights, there's just a lot of implicit leverage you get. They get seats at the AGI table. I don't know why we're giving authoritarian dictatorships the seat at the AGI table.

Dwarkesh Patel
There's going to be a lot of compute in the Middle East if these deals go through.

First of all, who is it? Is it just every single Big Tech company trying to figure it out over there?

Leopold Aschenbrenner
It’s not everybody, some.

Dwarkesh Patel
There are reports, I think Microsoft. We'll get into it.

So say the UAE gets a bunch of compute because we're building the clusters there. Let's say they have 25% of the compute. Why does a compute ratio matter? If it's about them being able to kick off the intelligence explosion, isn't it just some threshold where you have 100 million AI researchers or you don't?

Leopold Aschenbrenner
You can do a lot with 33 million extremely smart scientists. That might be enough to build the crazy bio weapons. Then you're in a situation where they stole the weights and they seized the compute.

Now they can make these crazy new WMDs that will be possible with superintelligence. Now you've just proliferated the stuff that'll be really powerful. Also, 3x on compute isn't actually that much.

The riskiest situation is if we're in some sort of really neck and neck, feverish international struggle. Say we're really close with the CCP and we're months apart. The situation we want to be in — and could be in if we play our cards right — is a little bit more like the US building the atomic bomb versus the German project years behind. If we have that, we just have so much more wiggle room to get safety right.

We're going to be building these crazy new WMDs that completely undermine nuclear deterrence. That's so much easier to deal with if you don't have somebody right on your tails and you have to go at maximum speed. You have no wiggle room. You're worried that at any time they can overtake you.

They can also just try to outbuild you. They might literally win. China might literally win if they can steal the weights, because they can outbuild you. They may have less caution, both good and bad caution in terms of whatever unreasonable regulations we have.

If you're in this really tight race, this sort of feverish struggle, that's when there's the greatest peril of self-destruction.

Dwarkesh Patel
Presumably the companies that are trying to build clusters in the Middle East realize this. Is it just that it's impossible to do this in America? If you want American companies to do this at all, do you have to do it in the Middle East or not at all? Then you just have China build a Three Gorges Dam cluster.

Leopold Aschenbrenner
There's a few reasons. People aren't thinking about this as the AGI superintelligence cluster. They're just like, "Ah, cool. Clusters for my ChatGPT."

Dwarkesh Patel
If you’re doing ones for inference, presumably you could spread them out across the country or something. The ones they’re building, they’re going to do one training run in a single thing they’re building.

Leopold Aschenbrenner
It’s just hard to distinguish between inference and training compute. People can claim it’s inference compute, but they might realize that actually this is going to be useful for training compute too.

Dwarkesh Patel
Because of synthetic data and things like that?

Leopold Aschenbrenner
RL looks a lot like inference, for example. Or you just end up connecting them in time. It's a lot like raw materials. It's like placing your uranium refinement facilities there.

So there are a few reasons. One, they don't think about this as the AGI cluster. Another is just that there’s easy money coming from the Middle East.

Another one is that some people think that you can't do it in the US. We actually face a real system competition here. Some people think that only autocracies that can do this with top-down mobilization of industrial capacity and the power to get stuff done fast.

Again, this is the sort of thing we haven't faced in a while. But during the Cold War, there was this intense system competition. East vs. West Germany was this. It was West Germany as liberal democratic capitalism vs. state-planned communism.

Now it's obvious that the free world would win. But even as late as 1961, Paul Samuelson was predicting that the Soviet Union would outgrow the United States because they were able to mobilize industry better.

So there are some people who shitpost about loving America, but then in private they're betting against America. They're betting against the liberal order. Basically, it's just a bad bet. This stuff is really possible in the US.

To make it possible in the US, to some degree we have to get our act together. There are basically two paths to doing it in the US. One is you just have to be willing to do natural gas. There's ample natural gas. You put your cluster in West Texas. You put it in southwest Pennsylvania by the Marcellus Shale. The 10 GW cluster is super easy. The 100 GW cluster is also pretty doable. I think natural gas production in the United States has almost doubled in a decade. You do that one more time over the next seven years, you could power multiple trillion-dollar data centers.

The issue there is that a lot of people made these climate commitments, not just the government. It's actually the private companies themselves, Microsoft, Amazon, etc., that have made these climate commitments. So they won't do natural gas. I admire the climate commitments, but at some point the national interest and national security is more important.

The other path is doing green energy megaprojects. You do solar and batteries and SMRs and geothermal. If we want to do that, there needs to be a broad deregulatory push. You can't have permitting take a decade. You have to reform FERC. You have to have blanket NEPA exemptions for this stuff.

There are inane state-level regulations. You can build the solar panels and batteries next to your data center, but it'll still take years because you actually have to hook it up to the state electrical grid. You have to use governmental powers to create rights of way to have multiple clusters and connect them and have the cables.

Ideally we do both. Ideally we do natural gas and the broader deregulatory green agenda. We have to do at least one. Then this stuff is possible in the United States.

Dwarkesh Patel
Before the conversation I was reading a good book about World War II industrial mobilization in the United States called Freedom's Forge. I'm thinking back on that period, especially in the context of reading Patrick Collison's Fast and the progress study stuff. There's this narrative out there that we had state capacity back then and people just got shit done but that now it's a clusterfuck.

Leopold Aschenbrenner
It wasn't at all the case!

Dwarkesh Patel
It was really interesting. You had people from the Detroit auto industry side, like William Knudsen, who were running mobilization for the United States. They were extremely competent. At the same time you had labor organization and agitation, which is very analogous to the climate change pledges and concerns we have today.

They would literally have these strikes, into 1941, costing millions of man-hours worth of time when we're trying to make tens of thousands of planes a month. They would just debilitate factories for trivial concessions from capital that were pennies on the dollar.

There were concerns that the auto companies were trying to use the pretext of a potential war to prevent paying labor the money it deserves. So with what climate change is today, you might think, "ah, America's fucked. We're not going to be able to build this shit if you look at NEPA or something," I didn't realize how debilitating labor was in World War II.

Leopold Aschenbrenner
It wasn' just that. Before 1939, the American military was in total shambles. You read about it and it reads a little bit like the German military today. Military expenditures were I think less than 2% of GDP. All the European countries had gone, even in peacetime, above 10% of GDP.

It was rapid mobilization starting from nothing. We were making no planes. There were no military contracts. Everything had been starved during the Great Depression. But there was this latent capacity. At some point the United States got its act together.

This applies the other way around too with China. Sometimes people count them out a little bit with the export controls and so on. They're able to make 7-nanometer chips now. There's a question of how many they could make. There's at least a possibility that they're going to mature that ability and make a lot of 7-nanometer chips.

There's a lot of latent industrial capacity in China. They are able to build a lot of power fast. Maybe that isn't activated for AI yet. At some point, the same way the United States and a lot of people in the US government are going to wake up, the CCP is going to wake up.

Dwarkesh Patel
Companies realize that scaling is a thing. Obviously their whole plans are contingent on scaling. So they understand that in 2028 we're going to be building 10 GW data centers.

At that point, the people who can keep up are Big Tech, potentially at the edge of their capabilities, sovereign wealth fund-funded things, and also major countries like America and China. What's their plan? With the AI labs, what's their plan given this landscape? Do they not want the leverage of being in the United States?

Leopold Aschenbrenner
The Middle East does offer capital, but America has plenty of capital. We have trillion-dollar companies. What are these Middle Eastern states? They're kind of like trillion-dollar oil companies. We have trillion-dollar companies and very deep financial markets. Microsoft could issue hundreds of billions of dollars of bonds and they can pay for these clusters.

Another argument being made, which is worth taking seriously, is that if we don't work with the UAE or with these Middle Eastern countries, they're just going to go to China. They're going to build data centers and pour money into AI regardless. If we don't work with them, they'll just support China.

There's some merit to the argument in the sense that we should be doing benefit-sharing with them. On the road to AGI, there should be two tiers of coalitions. There should be a narrow coalition of democracies that's developing AGI. Then there should be a broader coalition of other countries, including dictatorships, and we should offer them some of the benefits of AI.

If the UAE wants to use AI products, run Meta recommendation engines, or run the last-generation models, that's fine. By default, they just wouldn't have had this seat at the AGI table. So they have some money, but a lot of people have money.

The only reason they're getting this seat at the AGI table and giving these dictators this leverage over this extremely important national security technology, is because we're getting them excited and offering it to them.

Dwarkesh Patel
Who specifically is doing this? Who are the companies who are going there to fundraise?

Leopold Aschenbrenner
It’s been reported that Sam Altman is trying to raise $7 trillion or whatever for a chip project. It's unclear how many of the clusters will be there, but definitely stuff is happening.

There’s another reason I'm a little suspicious of this argument that if the US doesn't work with them, they'll go to China. I've heard from multiple people — not from my time at OpenAI, and I haven't seen the memo — that at some point several years ago, OpenAI leadership had laid out a plan to fund and sell AGI by starting a bidding war between the governments of the United States, China, and Russia.

It's surprising to me that they're willing to sell AGI to the Chinese and Russian governments. There's also something that feels eerily familiar about starting this bidding war and then playing them off each other, saying, "well, if you don't do this, China will do it."

Dwarkesh Patel
Interesting. That's pretty fucked up.

Suppose you're right. We ended up in this place because, as one of our friends put it, the Middle East has billions or trillions of dollars up for persuasion like no other place in the world.

Leopold Aschenbrenner
With little accountability. There's no Microsoft board. It's only the dictator.

Dwarkesh Patel
Let's say you're right, that you shouldn't have gotten them excited about AGI in the first place. Now we're in a place where they are excited about AGI and they're like, "fuck, we want to have GPT-5 while you're going to be off building superintelligence. This Atoms for Peace thing doesn't work for us." If you're in this place, don't they already have the leverage?

Leopold Aschenbrenner
The UAE on its own is not competitive. They're already export-controlled. You're not supposed to ship Nvidia chips over there. It's not like they have any of the leading AI labs. They have money, but it's hard to just translate money into progress.

Dwarkesh Patel
But I want to go back to other things you've been saying in laying out your vision. There's this almost industrial process of putting in the compute and algorithms, adding that up, and getting AGI on the other end. If it's something more like that, then the case for somebody being able to catch up rapidly seems more compelling than if it's some bespoke…

Leopold Aschenbrenner
Well, if they can steal the algorithms and if they can steal the weights, that's really important.

Dwarkesh Patel
How easy would it be for an actor to steal the things that are not the trivial released things, like Scarlett Johansson's voice, but the RL things we're talking about, the unhobblings?

Leopold Aschenbrenner
It's all extremely easy. They don't make the claim that it's hard. DeepMind put out their Frontier Safety Framework and they lay out security levels, zero to four. Four is resistant to state activity. They say, we're at level zero. Just recently, there was an indictment of a guy who stole a bunch of really important AI code and went to China with it. All he had to do to steal the code was copy it, put it into Apple Notes, and export it as a PDF. That got past their monitoring.

Google has the best security of any of the AI labs probably, because they have the Google infrastructure. I would think of the security of a startup. What does security of a startup look like? It's not that good. It's easy to steal.

Dwarkesh Patel
Even if that's the case, a lot of your post is making the argument for why we are going to get the intelligence explosion. If we have somebody with the intuition of an Alec Radford to come up with all these ideas, that intuition is extremely valuable and you can scale that up.

If it's just intuition, then that's not going to be just in the code, right? Also because of export controls, these countries are going to have slightly different hardware. You're going to have to make different trade-offs and probably rewrite things to be compatible with that.

Is it just a matter of getting the right pen drive and plugging it into the gigawatt data center next to the Three Gorges Dam and then you're off to the races?

Leopold Aschenbrenner
There are a few different things, right? One threat model is just them stealing the weights themselves. The weights one is particularly insane because they can just steal the literal end product — just make a replica of the atomic bomb — and then they're ready to go. That one is extremely important around the time we have AGI and superintelligence because China can build a big cluster by default. We'd have a big lead because we have the better scientists, but if we make the superintelligence and they just steal it, they're off to the races.

Weights are a little bit less important right now because who cares if they steal the GPT-4 weights. We still have to get started on weight security now because if we think there's AGI by 2027, this stuff is going to take a while. It's not just going to be like, "oh, we do some access control." If you actually want to be resistant to Chinese espionage, it needs to be much more intense.

The thing that people aren't paying enough attention to is the secrets. The compute stuff is sexy, but people underrate the secrets. The half an order of magnitude a year is just by default, sort of algorithmic progress. That's huge. If we have a few years of lead, by default, that's a 10-30x, 100x bigger cluster, if we protect them.

There's this additional layer of the data wall. We have to get through the data wall. That means we actually have to figure out some sort of basic new paradigm. So it's the “AlphaGo step two.” “AlphaGo step one” learns from human imitation. “AlphaGo step two” is the kind of self-play RL thing that everyone's working on right now. Maybe we're going to crack it. If China can't steal that, then they're stuck. If they can steal it, they're off to the races.

Dwarkesh Patel
Whatever that thing is, can I literally write it down on the back of a napkin? If it's that easy, then why is it so hard for them to figure it out? If it's more about the intuitions, then don't you just have to hire Alec Radford? What are you copying down?

Leopold Aschenbrenner
There are a few layers to this. At the top is the fundamental approach. On pre-training it might be unsupervised learning, next token prediction, training on the entire Internet. You actually get a lot of juice out of that already. That one's very quick to communicate.

Then there's a lot of details that matter, and you were talking about this earlier. It's probably going to be somewhat obvious in retrospect, or there's going to be some not too complicated thing that'll work, but there's going to be a lot of details to get that.

Dwarkesh Patel
If that's true, then again, why do we think that getting state-level security in these startups will prevent China from catching up? It’s just like, "Oh, we know some sort of self-play RL will be required to get past the data wall."

It's going to be solved by 2027, right? It's not that hard.

Leopold Aschenbrenner
The US, and the leading labs in the United States, have this huge lead. By default, China actually has some good LLMs because they're just using open source code, like Llama. People really underrate both the divergence on algorithmic progress and the lead the US would have by default because all this stuff was published until recently.

Look at Chinchilla Scaling laws, MoE papers, transformers. All that stuff was published. That's why open source is good and why China can make some good models. Now, they're not publishing it anymore. If we actually kept it secret, it would be a huge edge.

To your point about tacit knowledge and Alec Radford, there's another layer at the bottom that is something about large-scale engineering work to make these big training runs work. That is a little bit more like tacit knowledge, but China will be able to figure that out. It's engineering schlep, and they're going to figure out how to do it.

Dwarkesh Patel
Why can't they figure that out, but not how to get the RL thing working?

Leopold Aschenbrenner
I don't know. Germany during World War II went down the wrong path with heavy water. There's an amazing anecdote in The Making of the Atomic Bomb about this.

Secrecy was one of the most contentious issues early on. Leo Szilard really thought a nuclear chain reaction and an atomic bomb were possible. He went around saying, "this is going to be of enormous strategic and military importance." A lot of people didn't believe it or thought, "maybe this is possible, but I'm going to act as though it's not, and science should be open."

In the early days, there had been some incorrect measurements made on graphite as a moderator. Germany thought graphite wasn't going to work, so they had to do heavy water. But then Enrico Fermi made new measurements indicating that graphite would work. This was really important.

Szilard assaulted Fermi with another secrecy appeal and Fermi was pissed off, throwing a temper tantrum. He thought it was absurd, saying, "Come on. This is crazy." But Szilard persisted, and they roped in another guy, George Pegram. In the end, Fermi didn't publish it.

That was just in time. Fermi not publishing meant that the Nazis didn't figure out graphite would work. They went down the path of heavy water, which was the wrong path. This is a key reason why the German project didn't work out. They were way behind.

We face a similar situation now. Are we just going to instantly leak how to get past the data wall and what the next paradigm is? Or are we not?

Dwarkesh Patel
The reason this would matter is if being one year ahead would be a huge advantage. In the world where you deploy AI over time they're just going to catch up anyway.

I interviewed Richard Rhodes, the guy who wrote The Making of the Atomic Bomb. One of the anecdotes he had was when the Soviets realized America had the bomb. Obviously, we dropped it in Japan.

Lavrentiy Beria — the guy who ran the NKVD, a famously ruthless and evil guy — goes to the Soviet scientist who was running their version of the Manhattan Project. He says, "comrade, you will get us the American bomb." The guy says, "well, listen, their implosion device actually is not optimal. We should make it a different way." Beria says, "no, you will get us the American bomb, or your family will be camp dust."

The thing that's relevant about that anecdote is that the Soviets would have had a better bomb if they hadn't copied the American design, at least initially. That suggests something about history, not just for the Manhattan Project. There's often this pattern of parallel invention because the tech tree implies that a certain thing is next — in this case, a self-play RL — and people work on that and are going to figure it out around the same time. There's not going to be that much gap in who gets it first.

Famously, a bunch of people invented the light bulb around the same time. Is it the case that it might be true but the one year or six months makes the difference?

Leopold Aschenbrenner
Two years makes all the difference.

Dwarkesh Patel
I don't know if it'll be two years though.

Leopold Aschenbrenner
If we lock down the labs, we have much better scientists. We're way ahead. It would be two years. Even six months, a year, would make a huge difference. This gets back to the intelligence explosion dynamics. A year might be the difference between a system that's sort of human-level and a system that is vastly superhuman. It might be like five OOMs.

Look at the current pace. Three years ago, on the math benchmark — these are really difficult high school competition math problems — we were at a few percent, we couldn't solve anything. Now it's solved. That was at the normal pace of AI progress. You didn't have a billion superintelligent researchers.

A year is a huge difference, particularly after superintelligence. Once this is applied to many elements of R&D, you get an industrial explosion with robots and other advanced technologies. A couple of years might yield decades worth of progress. Again, it’s like the technological lead the U.S. had in the first Gulf War, when the 20-30 years of technological lead proved totally decisive. It really matters.

Here’s another reason it really matters. Suppose they steal the weights, suppose they steal the algorithms, and they're close on our tails. Suppose we still pull out ahead. We're a little bit faster and we're three months ahead.

The world in which we're really neck and neck, we only have a three-month lead, is incredibly dangerous. We're in this feverish struggle where if they get ahead, they get to dominate, maybe they get a decisive advantage. They're building clusters like crazy. They're willing to throw all caution to the wind. We have to keep up.

There are crazy new WMDs popping up. Then we're going to be in the situation where it's crazy new military technology, crazy new WMDs, deterrence, mutually assured destruction keeps changing every few weeks. It's a completely unstable, volatile situation that is incredibly dangerous.

So you have to look at it from the point of view that these technologies are dangerous, from the alignment point of view. It might be really important during the intelligence explosion to have a six-month wiggle room to be like, “look, we're going to dedicate more compute to alignment during this period because we have to get it right. We're feeling uneasy about how it's going.”

One of the most important inputs to whether we will destroy ourselves or whether we will get through this incredibly crazy period is whether we have that buffer.

Dwarkesh Patel
Before we go further, it's very much worth noting that almost nobody I talk to thinks about the geopolitical implications of AI. I have some object-level disagreements that we'll get into, things I want to iron out. I may not disagree in the end.

The basic premise is that if you keep scaling, if people realize that this is where intelligence is headed, it's not just going to be the same old world. It won't just be about what model we're deploying tomorrow or what the latest thing is. People on Twitter are like, "oh, GPT-4 is going to shake your expectations" or whatever.

COVID is really interesting because when March 2020 hit, it became clear to the world — presidents, CEOs, media, the average person — that there are other things happening in the world right now but the main thing we as a world are dealing with right now is COVID.

Leopold Aschenbrenner
Soon it will be AGI. This is the quiet period. Maybe you want to go on vacation. Maybe now is the last time you can have some kids. My girlfriend sometimes complains when I'm off doing work that I don't spend enough time with her. She threatens to replace me with GPT-6 or whatever. I'm like, "GPT-6 will also be too busy doing AI research."

Dwarkesh Patel
Why aren't other people talking about national security?

Leopold Aschenbrenner
I made this mistake with COVID. In February of 2020, I thought it was going to sweep the world and all the hospitals would collapse. It would be crazy, and then it'd be over. A lot of people thought this kind of thing at the beginning of COVID. They shut down their office for a month or whatever.

The thing I just really didn't price in was societal reaction. Within weeks, Congress spent over 10% of GDP on COVID measures. The entire country was shut down. It was crazy. I didn't sufficiently price it in with COVID.

Why do people underrate it? Being in the trenches actually gives you a less clear picture of the trend lines. You don't have to zoom out that much, only a few years.

When you're in the trenches, you're trying to get the next model to work. There's always something that's hard. You might underrate algorithmic progress because you're like, "ah, things are hard right now," or "data wall" or whatever. When you zoom out just a few years and count up how much algorithmic progress was made, it's enormous.

People also just don’t think about this stuff. Smart people really underrate espionage. Part of the security issue is that people don't realize how intense state-level espionage can be. This Israeli company had software that could just zero-click hack any iPhone. They just put in your number and it was a straight download of everything. The United States infiltrated an air-gapped atomic weapons program. Wild.

Dwarkesh Patel
Are you talking about Stuxnet?

Leopold Aschenbrenner
Yeah. Intelligence agencies have stockpiles of zero-days. When things get really hot, maybe we'll send special forces to go to the data center or something. China does this. They threaten people's families. They’re like, “If you don't cooperate - if you don't give us the intel…”

There's a good book along the lines of The Gulag Archipelago called Inside the Aquarium, which is by a Soviet GRU defector. GRU was military intelligence. Ilya recommended this book to me. When I read it, I was shocked at the intensity of state-level espionage.

The whole book was about how they go to these European countries and try and recruit people to get the technology. Here’s one anecdote. This eventual defector, he's being trained at the GRU spy academy. To graduate from the spy academy before being sent abroad, you had to pass a test to show that you can do this.

The test was recruiting a Soviet scientist in Moscow to give you information, like you would do in a foreign country. Of course, for whomever you recruited, the penalty for giving away secret information was death. So to graduate from the GRU spy academy, you had to condemn a countryman to death. States do this stuff.

Dwarkesh Patel
I started reading the book because you mentioned it in the series. I was wondering about the fact that you use this anecdote. Then you're like, "A book recommended by Ilya." Is this some sort of Easter egg? We'll leave that as an exercise for the reader.

Leopold Aschenbrenner
The beatings will continue until morale improves.

Dwarkesh Patel
Suppose we live in a world where these secrets are locked down, but China realizes this progress is happening in America.

Leopold Aschenbrenner
The secrets probably won't be locked down. We're probably going to live in the bad world. It's going to be really bad.

Dwarkesh Patel
Why are you so confident they won't be locked down?

Leopold Aschenbrenner
I'm not confident they won't be locked down, but it's just not happening.

Dwarkesh Patel
Let's say tomorrow, the lab leaders get the message. How hard is it? What do they have to do? Do they get more security guards? Do they air-gap? What do they do?

Leopold Aschenbrenner
People have two reactions: "We're already secure." We're not.

Then there's fatalism: "It's impossible."

You need to stay ahead of the curve of how AGI-pilled the CCP is. Right now, you've got to be resistant to normal economic espionage. They're not. I probably wouldn't be talking about this stuff if the labs were. I wouldn't want to wake up the CCP more. But this stuff is really trivial for them to do right now.

So, they're not resistant to that. It would be possible for a private company to be resistant to it. Both of us have friends in the quantitative trading world. Those secrets are shaped similarly where if I got on a call for an hour with somebody from a competitor firm, most of our alpha would be gone.

Dwarkesh Patel
You're going to worry about that pretty soon.

Leopold Aschenbrenner
All the alpha could be gone but in fact, their alpha often persists for many years and decades. So this doesn't seem to happen. There's a lot you could do if you went from current startup security to good private sector security: hedge funds, the way Google treats customer data or whatever. That'd be good right now.

The issue is that basically the CCP will also get more AGI-pilled. At some point, we're going to face the full force of the Ministry of State Security. You're talking about smart people underrating espionage and the insane capabilities of states. This stuff is wild. There are papers about how you can find out the location of where you are in a video game map just from sounds. States can do a lot with electromagnetic emanations.

At some point, you have to be working from a SCIF. Your cluster needs to be air-gapped and basically be a military base. You need to have intense security clearance procedures for employees. All this shit is monitored. They basically have security guards. You can't use any other dependencies. It's all got to be intensely vetted. All your hardware has to be intensely vetted.

If they actually really face the full force of state-level espionage, this isn't really the thing private companies can do empirically. Microsoft recently had executives' emails hacked by Russian hackers, and government emails they've hosted hacked by government actors. Also, there's just a lot of stuff that only the people behind the security clearances know and only they deal with.

To actually resist the full force of espionage, you're going to need the government. We could do it by always being ahead of the curve. I think we're just going to always be behind the curve, unless we get a sort of government project.

Dwarkesh Patel
Going back to the naive perspective, we're very much coming at this from, “There's going to be a race and the CCP, we must win.” Listen, I understand bad people are in charge of the Chinese government, with the CCP and everything.

I want to step back to a sort of galactic perspective. Humanity is developing AGI. Do we want to come at this from the perspective of "we need to beat China"? To our superintelligent Jupiter brain descendants, China will be some distant memory that they have, America too.

Shouldn't it be more, as an initial approach, just going to them like, “listen, this is superintelligence. We come from a cooperative perspective.” Why immediately rush into it from a hawkish, competitive perspective?

Leopold Aschenbrenner
A lot of the stuff I talk about in the series is primarily descriptive. On the China stuff, in some ideal world, it's just all merry-go-round and cooperation. Again, people wake up to AGI. The issue in particular is, can we make a deal? Can we make an international treaty? It really relates to the stability of international arms control agreements.

We did very successful arms control on nuclear weapons in the 1980s. The reason it was successful is because the new equilibrium was stable. You go down from 60,000 nukes to 10,000 nukes or whatever. When you have 10,000 nukes, breakout basically doesn't matter that much.

Suppose the other guy now tried to make 20,000 nukes. Who cares? It's still mutually assured destruction. Suppose a rogue state went from zero nukes to one nuke. Who cares? We still have way more nukes than you. It's still not ideal for destabilization.

It'd be very different if the arms control agreement had been zero nukes. At zero nukes, you just need one rogue state to make one nuke and the whole thing is destabilized. Breakout is very easy. Your adversary state starts making nukes.

When you're going to very low levels of arms or when you're in a very dynamic technological situation, arms control is really tough because breakout is easy. There are some other stories about this in the 1920s and 1930s. All the European states had disarmed.

Germany did this kind of crash program to build the Luftwaffe. That was able to massively destabilize things because they were the first. They were able to pretty easily build a modern air force because the others didn't really have one. That really destabilized things.

The issue with AGI and superintelligence is the explosiveness of it. If you have an intelligence explosion, you're able to go from AGI to superintelligence. That superintelligence is decisive because you'll developed some crazy WMD or you'll have some super hacking ability that lets you completely deactivate the enemy arsenal. Suppose you're trying to put in a break. We're both going to cooperate. We're going to go slower on the cusp of AGI.

There is going to be such an enormous incentive to race ahead, to break out. We're just going to do the intelligence explosion. If we can get three months ahead, we win. That makes any sort of arms control agreement very unstable in a close situation.

Dwarkesh Patel
That's really interesting. This is very analogous to a debate I had with Rhodes on the podcast where he argued for nuclear disarmament. If some country tried to break out and started developing nuclear weapons, the six months you would get is enough to get international consensus and invade the country and prevent them from getting nukes. I thought that was not stable equilibrium.

On this, maybe it's a bit easier because you have AGI and so you can monitor the other person's cluster or something. You can see the data centers from space. You can see the energy draw they're getting. As you were saying, there are a lot of ways to get information from an environment if you're really dedicated. Also, unlike nukes, the data centers are fixed. Obviously, you have nukes in submarines, planes, bunkers, mountains, etc. You can have them so many different places. A 100 GW data center, we can blow that shit up if we're concerned. We can just use a cruise missile or something. That's very vulnerable.

Leopold Aschenbrenner
That gets to the insane vulnerability and the volatility of this period, post-superintelligence. You have the intelligence explosion. You have these vastly superhuman things on your cluster. You haven't done the industrial explosion yet. You don't have your robots yet. You haven't covered the desert in robot factories yet.

That is this crazy moment. Say the United States is ahead. The CCP is somewhat behind. There's actually an enormous incentive for first strike, if they can take out your data center. They know you're about to have this command, a decisive lead. They know if they can just take out this data center, then they can stop it. They might get desperate.

We're going to get into a position that's going to be pretty hard to defend early on. We're basically going to be in a position where we're protecting data centers with the threat of nuclear retaliation. Maybe it sounds kind of crazy.

Dwarkesh Patel
Is this the inverse of the Eliezer…?

Leopold Aschenbrenner
Nuclear deterrence for data centers. This is Berlin in the late 1950s, early 1960s. Both Eisenhower and Kennedy multiple times made the threat of full-on nuclear war against the Soviets if they tried to encroach on West Berlin.

It's sort of insane. It's kind of insane that that went well. Basically, that's going to be the only option for the data centers. It's a terrible option. This whole scheme is terrible. Being in neck and neck race at this point is terrible.

I have some uncertainty on how easy that decisive advantage will be. I'm pretty confident that if you have superintelligence, you have two years, you have the robots, you're able to get that 30-year lead. Then you're in this Gulf War 1 situation. You have your millions or billions of mosquito-sized drones that can just take it out. There's even a possibility you can get a decisive advantage earlier.

There are these stories about colonization in the 1500s where a few hundred Spaniards were able to topple the Aztec Empire, a couple of other empires as well. Each of these had a few million people. It was not a godlike technological advantage. It was some technological advantage. It was some amount of disease and cunning strategic play.

There's a possibility that even early on — when you haven't gone through the full industrial explosion yet — you have superintelligence, but you're able to manipulate the opposing generals, claiming you're allying with them. Then you have some crazy new bioweapons. Maybe there's even some way to pretty easily get a paradigm that deactivates enemy nukes. This stuff could get pretty wild.

Here's what we should do. I really don't want this volatile period. A deal with China would be nice. It's going to be really tough if you're in this unstable equilibrium. We want to get in a position where it is clear that the United States, a coalition of democratic allies, will win. It is clear to the United States, it is clear to China. That will require having locked down the secrets, having built the 100 gigawatt cluster in the United States, having done the natural gas and doing what's necessary.

When it is clear that the democratic coalition is well ahead, you go to China and offer them a deal. China will know we're going to win. They're very scared of what's going to happen. We're going to know we're going to win, but we're also very scared of what's going to happen because we really want to avoid this kind of breakneck race right at the end. Things could really go awry.

We offer them a deal. There's an incentive to come to the table. There's a more stable arrangement you can do. It's an Atoms for Peace arrangement. We're like, "look, we're going to respect you. We're not going to use superintelligence against you. You can do what you want. You're going to get your slice of the galaxy.

We're going to benefit-share with you. We're going to have some compute agreement where there's some ratio of compute that you're allowed to have, enforced with opposing AIs or whatever. We're just not going to do this volatile WMD arms race to the death.

It's a new world order that's US-led, democracy-led, but respects China and lets them do what they want.

Dwarkesh Patel
There's so much there. On the galaxies thing, there's a funny anecdote. I kind of want to tell it. We were at an event. I'm respecting Chatham House rules here. I'm not revealing anything about it. Leopold was talking to somebody influential. Afterwards, that person told the group, "Leopold told me he's not going to spend any money on consumption until he's ready to buy galaxies."

The guy goes, "I honestly don't know if he meant galaxies like the brand of private plane Galaxy or physical galaxies." There was an actual debate. He went away to the restroom. There was an actual debate among influential people about whether he meant Galaxys. Others who knew you better were like, "No, he means galaxies."

Leopold Aschenbrenner
I meant the galaxies. There are two ways to buy the galaxies. At some point, post-superintelligence, there's some crazy...

Dwarkesh Patel
I'm laughing my ass off, not even saying anything. We were having this debate. Leopold comes back. Someone says, "oh, Leopold, we're having this debate about whether you meant you want to buy the Galaxy, or you want to buy the other thing." Leopold assumes they must mean not the private plane Galaxy vs. the actual galaxy, but whether he wants to buy the property rights of the galaxy or actually just send out the probes right now.

Leopold Aschenbrenner
Exactly.

Dwarkesh Patel
Alright, back to China. There's a whole bunch of things I could ask about that plan and whether you're going to get a credible promise to get some part of galaxies.

Leopold Aschenbrenner
You'll have AIs to help you enforce stuff.

Dwarkesh Patel
Sure, we'll leave that aside. That's a different rabbit hole. The thing I want to ask is...

Leopold Aschenbrenner
The only way this is possible is if we lock it down. If we don't lock it down, we are in this fever struggle. Greatest peril mankind will have ever seen.

Dwarkesh Patel
During this period, they don't really understand how this AI governance is going to work, whether they're going to check, whether we're going to adjugate the galaxies. The data centers can't be built underground. They have to be above ground. Taiwan is right off the coast of China. The US needs the chips from there.

Why isn't China just going to invade? Worst case scenario for them is the US wins the superintelligence, which we're on track to do anyway. Wouldn't this instigate them to either invade Taiwan or blow up the data center in Arizona or something like that?

Leopold Aschenbrenner
You talked about the data center. You'd probably have to threaten nuclear retaliation to protect that. They might just blow it up. There are also ways they can do it without attribution.

Dwarkesh Patel
Stuxnet.

Leopold Aschenbrenner
Stuxnet, yeah. We'll talk about later, but we need to be working on a Stuxnet for the Chinese project. I talk about AGI by 2027 or whatever. On Taiwan, do you know about the terrible twenties?

Dwarkesh Patel
No.

Leopold Aschenbrenner
In Taiwan watcher circles, people often talk about the late 2020s as the maximum period of risk for Taiwan. Military modernization cycles and extreme fiscal tightening on the US military budget over the last decade or two have meant that we're in a trough by the late twenties in terms of overall naval capacity.

That's when China is saying they want to be ready. It's already kind of a parallel timeline there. Yeah, it looks appealing to invade Taiwan. Maybe not because of the remote cut off chips, which deactivates the machines. But imagine if during the Cold War, all of the world's uranium deposits had been in Berlin. Berlin already almost caused a nuclear war multiple times. God help us all.

Dwarkesh Patel
Leslie Groves actually had a plan after the war that America would go around the world getting the rights to every single uranium deposit because they didn't realize how much uranium there was in the world. They thought this was feasible. They didn't realize, of course, that there were huge deposits in the Soviet Union itself.

Leopold Aschenbrenner
East Germany, too. A lot of East German workers got screwed and got cancer.

Dwarkesh Patel
The framing we've been assuming — I'm not sure I buy it yet—is that the United States has this leverage. This is our data center. China is the competitor right now. Obviously, that's not the way things are progressing. Private companies control these AIs. They're deploying them. It's a market-based thing. Why will it be the case that the United States has this leverage or is doing this thing versus China doing this thing?

Leopold Aschenbrenner
There are descriptive and prescriptive claims, or normative and positive claims. The main thing I'm trying to say is, at these SF parties, people talk about AGI and always focus on private AI labs. I want to challenge that assumption.

It seems likely to me, for reasons we've discussed, that the national security state will get involved. There are many ways this could look: nationalization, a public-private partnership, a defense contractor-like relationship, or a government project that absorbs all the people. There's a spectrum, but people vastly underrate the chances of this looking like a government project.

When we have literal superintelligence on our cluster — with a billion superintelligent scientists who can hack everything and Stuxnet the Chinese data centers, and build robo armies — you really think it'll be a private company ? The government would be like, "oh, my God, what is going on?"

Dwarkesh Patel
Suppose there's no China. Suppose there are countries like Iran and North Korea that theoretically could achieve superintelligence, but they're not on our heels. In that world, are you advocating for a national project or do you prefer the private path forward?

Leopold Aschenbrenner
Two responses to this. One is, you still have Russia and other countries.

You need Russia-proof security. You can't let Russia steal all your stuff. Their clusters may not be as big, but they can still make crazy bioweapons and mosquito-sized drone swarms.

The security component is a large part of the project because there's no other way to prevent this from instantly proliferating to everyone. You still have to deal with Russia, Iran, and North Korea. Saudi and Iran will try to get it to screw each other. Pakistan and India will try to get it to screw each other. There's enormous destabilization.

Still, I agree with you. If AGI had emerged in 2005, during unparalleled American hegemony, there would have been more scope for less government involvement. But as we discussed, that would have been a unique moment in history. In almost all other moments in history, there would have been a great power competitor.

Dwarkesh Patel
Let's get into this debate. My position is this. If you look at the people who were involved in the Manhattan Project, many of them regretted their participation. We can infer from this that we should start with a cautious approach to the nationalized ASI project.

Leopold Aschenbrenner
Did they regret their participation because of the project or because of the technology itself? People will regret it, but it's about the nature of the technology, not the project.

Dwarkesh Patel
They probably had a sense that different decisions would have been made if it wasn't a concerted effort that everyone agreed to participate in. If it wasn't in the context of a race to beat Germany and Japan, you might not develop it. That's the technology part.

Leopold Aschenbrenner
It's still going to be a weapon because of the destructive potential, the military potential. It's not because of the project. It's because of the technology. That will unfold regardless.

Imagine you go through the 20th century in a decade —

Dwarkesh Patel
Let's run that example. Suppose the 20th century was run through in one decade.

Do you think the technologies that happened during the 20th century shouldn't have been privatized? Should it have been a more concerted, government-led project?

Leopold Aschenbrenner
There's a history of dual-use technologies. AI will be dual-use in the same way. There will be lots of civilian uses of it. Like with nuclear energy, the government project developed the military angle of it and then worked with private companies. There was a flourishing of nuclear energy until the environmentalists stopped it.

Planes, like Boeing. Actually, the Manhattan Project wasn't the biggest defense R&D project during World War II. It was the B-29 bomber because they needed a bomber with a long enough range to reach Japan to destroy their cities. Boeing made the B-47, and the B-52 plane the US military uses today. They used that technology later on to build the 707.

Dwarkesh Patel
What does "later on" mean in this context? I get what it means after a war to privatize. But if the government has ASI…

Let me back up and explain my concern. You have this institution in our society with a monopoly on violence. We're going to give it access to ASI that's not broadly deployed. This maybe sounds silly, but we're going to go through higher levels of intelligence. Private companies will be required by regulation to increase their security. They'll still be private companies.

They'll deploy this and release AGI. Now McDonald's, JP Morgan, and some random startup will be more effective organizations because they have AGI workers. It'll be like the Industrial Revolution, where the benefits were widely diffused.

Backing up, what is it we're trying to do? Why do we want to win against China? We want to win because we don't want a top-down authoritarian system to win. If the way to beat that is for the most important technology for humanity to be controlled by a top-down government, what's the point?

Let's run our cards with privatization. That's how we get to the classic liberal, market-based system we want for the ASIs.

Leopold Aschenbrenner
Alright, there's a lot to talk about here. I'll start by looking at what the private world would look like. This is part of why there's no alternative. Then let's look at what the government project looks like, what checks and balances look like, and so on.

Let's start with the private world. A lot of people talk about open source. There's a misconception that AGI development will be a beautiful, decentralized thing, a giddy community of coders collaborating. That's not how it's going to look. It's a $100 billion or trillion-dollar cluster. Not many people will have it.

Right now, open source is good because people use the stuff that's published. They use the published algorithms, or, like Mistral, they leave DeepMind, take all the secrets, and replicate it.

That's not going to continue. People also say stuff like, "10^26 flops will be in my phone." No, it won't. Moore's Law is really slow. AI chips are getting better but the $100 billion computer won't cost $1,000 within your lifetime. So it's going to be like two or three big players in the private world.

You talk about the enormous power that superintelligence and the government will have. It's pretty plausible that in the alternative world one AI company will have that power. Say OpenAI has a six-month lead. You're talking about the most powerful weapon ever. You're making a radical bet on a private company CEO as the benevolent dictator.

Dwarkesh Patel
Not necessarily. Like any other thing that's privatized, we don't count on them being benevolent. Think of someone who manufactures industrial fertilizer. This person with this factory, if they went back to an ancient civilization, they could blow up Rome. They could probably blow up Washington, DC.

Leopold Aschenbrenner
Indeed.

Dwarkesh Patel
In your series, you talk about Tyler Cowen's phrase of "muddling through." Even with privatization, people underrate that there are a lot of private actors who control vital resources like the water supply.

We can count on cooperation and market-based incentives to maintain a balance of power. Sure, things are proceeding really fast. We have a lot of historical evidence that this works best.

Leopold Aschenbrenner
What do we do with nukes, right? We don't keep nukes in check by beefing up the Second Amendment so each state has its own nuclear arsenal. Dario and Sam don't have their own little arsenal.

No, it's institutions, constitutions, laws, and courts. I'm not sure this balance of power analogy holds. The government having the biggest guns was an enormous civilizational achievement, like Landfrieden in the Holy Roman Empire. If someone from the neighboring town committed a crime, you didn't start a battle between the towns. You took it to a court of the Holy Roman Empire. They decided it. It's a big achievement.

The key differences with the analogy about the industrial fertilizer are speed and offense-defense balance issues. It's like compressing the 20th century into a few years. That is incredibly scary because of the rapid advancement in destructive technology and military advancements.

You'd go from bayonets and horses to tank armies and fighter jets in a couple of years. In just a few more years you'd have nukes, ICBMs, and stealth. That speed creates an incredibly volatile and dangerous period. We have to make it through that, which will be incredibly challenging.

That's where a government project is necessary. If we can make it through that, the situation stabilizes. We don't face this imminent national security threat. Yes, there were WMDs that developed, but we've managed to create a stable offense-defense balance.

Bioweapons are a huge issue initially. An attacker can create 1000 different synthetic viruses and spread them. It's hard to defend against each. Maybe at some point, you figure out a universal defense against every possible virus, then you're in a stable situation again on the offense-defense balance. Or like with planes, you restrict certain capabilities that the private sector isn't allowed to have, then you can let the civilian uses run free.

Dwarkesh Patel
I’m skeptical of this.

Leopold Aschenbrenner
This is the other important thing. I talked about one company having all this power. It is unprecedented because the industrial fertilizer guy cannot overthrow the US government. It’s quite plausible that the AI company with superintelligence can.

Dwarkesh Patel
There would be multiple AI companies, right? I buy that one of them could be ahead.

Leopold Aschenbrenner
It’s not obvious that it’ll be multiple. If there’s a six-month lead, maybe there are two or three.

Dwarkesh Patel
I agree.

Leopold Aschenbrenner
If there are two or three, then it’s a crazy race between these companies. Demis and Sam would be like, "I don’t want to let the other one win." They’re both developing their nuclear arsenals and robots.

Come on. The government is not going to let these people do that. Is Dario going to be the one developing super hacking Stuxnet and deploying it against the Chinese data center?

The other issue is that if it’s two or three, it won’t just be two or three. It’ll be China, Russia, and North Korea too. In the private lab world, there’s no way they’ll have good enough security.

Dwarkesh Patel
We’re also assuming that if you nationalize it, especially in a world where this stuff is priced in by the CCP, you’ve got it nailed down. I’m not sure why we would expect that.

Leopold Aschenbrenner
The government’s the only one who does this stuff.

Dwarkesh Patel
If we don’t trust Sam or Dario to be benevolent dictators…

Leopold Aschenbrenner
Just corporate governance in general.

Dwarkesh Patel
Because you can cause a coup, the same capabilities are going to be true of the government project, right? The modal president in 2025, Donald Trump, will be the person versus you not trusting Sam or Dario to have these capabilities. I agree that if Sam and Dario have a one-year lead on ASI, in that world I'm concerned about privatization.

In that exact same world, I'm very concerned about Donald Trump having the capability. Potentially, if the takeoff is slower than anticipated, I prefer the private companies in that world. In no part of this matrix is it obviously true that the government-led project is better.

Leopold Aschenbrenner
Let's talk about the government project and checks and balances.

In some sense, my argument is a Burkean one. American checks and balances have held for over 200 years through crazy technological revolutions. The US military could kill every civilian in the United States.

Dwarkesh Patel
You're going to make that argument. The private-public balance of power has held for hundreds of years.

Leopold Aschenbrenner
But, why has it held? It's because the government has had the biggest guns. Never before has a single CEO or a random nonprofit board had the ability to launch nukes.

What is the track record of government checks and balances versus the track record of the private company checks and balances? Well the AI lab's first stress test went really badly.

Even worse in the private company world, it's two private companies and the CCP. They'll just instantly have all the tech. They probably won't have good enough internal control. It's not just the random CEO, but rogue employees who can use these superintelligences to do whatever they want.

Dwarkesh Patel
This won't be true of the government? Rogue employees won't exist on the project?

Leopold Aschenbrenner
The government has actual decades of experience and actually cares about this stuff. They deal with nukes and really powerful technology. This is the stuff that the national security state cares about.

Let's talk about government checks and balances a little bit. What are checks and balances in the government world? First, it's important to have some international coalition. I talked about these two tiers before. The inner tiers are modeled on the Quebec Agreement, Churchill and Roosevelt agreeing to pool efforts on nukes but not using them against each other, or anyone else without consent.

Bring in the UK with DeepMind, Southeast Asian states with the chip supply chain, and more NATO allies with talent and industrial resources. You have those checks and balances with more international countries at the table.

Separately, you have the second tier of coalitions, the Atoms for Peace thing. You go to countries including the UAE and make a deal similar to the NPT. They're not allowed to do crazy military stuff, but we'll share civilian applications. We'll help them and share the benefits, creating a new post-superintelligence world order.

US checks and balances: Congress will have to be involved to appropriate trillions of dollars. Ideally, Congress needs to confirm whoever's running this. You have Congress, different factions of the government, and the courts. I expect the First Amendment to remain really important.

This sounds crazy to people, but these institutions have withstood the test of time in a powerful way. This is why alignment is important. You program AIs to follow the constitution. The military works because generals are not allowed to follow unlawful or unconstitutional orders. You have the same thing for the AIs.

Dwarkesh Patel
So what's wrong with this argument. Maybe you have a point in a world with an extremely fast takeoff, one year from AGI to ASI.

Leopold Aschenbrenner
Then you have the years after ASI where you have this extraordinary explosion and technological progress.

Dwarkesh Patel
Maybe you have a point. We don't know. You have arguments for why that's a more likely world, but maybe that's not the world we live in.

In the other world, I'm very much on the side of ensuring these things are privately held. When you nationalize, that's a one-way function. You can't go back.

Why not wait until we have more evidence on which world we live in? Rushing nationalization might be a bad idea while we're uncertain. I'll let you respond to that first.

Leopold Aschenbrenner
I don't expect us to nationalize tomorrow. If anything I expect it to be like COVID,where it's kind of too late. Ideally, you nationalize early enough to lock stuff down. It'll probably be chaotic. You'll be trying to do a crash program to lock stuff down. It'll be kind of late. It'll be clear what's happening. We're not going to nationalize when it's not clear what's happening.

Dwarkesh Patel
The argument that these institutions have held up historically so well is flawed. They've actually almost broken a bunch of times.

Leopold Aschenbrenner
They've held up. They didn't break the first time they were tested.

Dwarkesh Patel
This is similar to the argument that some people make about nuclear war: we've had nukes for 80 years and have been fine, so the risk must be low. The answer to that is no. The risk is really high. We've avoided it because people have made a lot of effort to prevent it. Giving the government ASI without knowing the implications isn't making that effort.

Look at the base rate. America is very exceptional, not just in terms of avoiding dictatorship. Every other country in history has had a complete drawdown of wealth because of war, revolution, etc. America is very unique in not having had that.

We have to think of the historical base rate. We haven't thought about great power competition in the last 80 years, but it's significant. Dictatorship is the default state of mankind. Relying on institutions in an ASI world is fundamentally different. Right now, if the government tried to overthrow, it's much harder without ASI. There are people with AK-47s and AR-15s, making it harder.

Leopold Aschenbrenner
The government could crush the AR-15s.

Dwarkesh Patel
No, it would actually be pretty hard. It's the reason why Vietnam and Afghanistan were so hard.

Leopold Aschenbrenner
They could just nuke the whole country.

Dwarkesh Patel
I agree.

Leopold Aschenbrenner
They could. It's similar to the ASI.

Dwarkesh Patel
It's just easier if you have what you were talking.

Leopold Aschenbrenner
No, there are institutions, constitutions, legal restraints, courts, and checks and balances. The crazy bet is the bet on a private company CEO.

Dwarkesh Patel
Isn't the same thing true of nukes where we have institutional agreements about non-proliferation? We're still very concerned about those being broken and someone getting nukes. We stay up at night worrying about that situation.

Leopold Aschenbrenner
It's a precarious situation. ASI will a precarious situation as well. Given how precarious nukes are, we've done pretty well.

Dwarkesh Patel
What does privatization in this world even mean? What happens after?

Leopold Aschenbrenner
We're talking about whether the government project is good or not. I have very mixed feelings about this as well.

My primary argument is that if you're at the point where this thing has vastly superhuman capabilities — it can develop crazy bioweapons targeted to kill everyone but the Han Chinese, it can wipe out entire countries, it can build robo armies and drone swarms with mosquito-sized drones — the US national security state will be intimately involved.

The government project will look like a joint venture between cloud providers, labs, and the government. There is no world in which the government isn't involved in this crazy period. At the very least, intelligence agencies need to run security for these labs. They're already controlling access to everything.

If we're in a volatile international situation, initial applications will focus on stabilizing it. It'll suck. It's not what I want to use ASI for. Somehow we need to prevent proliferation of new WMDs, and maintain mutually assured destruction with North Korea, Russia, and China.

There's a broader spectrum than you're acknowledging. In a world with private labs, there will be heavy government involvement. What we're debating is the form of government involvement, but it will look more like the national security state than a startup, which is what it is right now.

Dwarkesh Patel
Something like that makes sense. I'd be very worried if it's like the Manhattan Project, where it's directly part of the US military. If it's more like needing to talk to Jake Sullivan before running the next training line…

Leopold Aschenbrenner
Is Lockheed Martin's Skunk Works part of the US military? They call the shots.

Dwarkesh Patel
I don't think that's great and I think that's bad if it happens with ASI. What's the scenario?

Leopold Aschenbrenner
What's the alternative? What's the alternative?

Dwarkesh Patel
It's closer to my end of the spectrum. You talk to Jake Sullivan before you can launch the next training cluster, but many companies are still going for it, and the government is involved in security.

Leopold Aschenbrenner
Is Dario launching the Stuxnet attack?

Dwarkesh Patel
What do you mean by launching?

Leopold Aschenbrenner
Dario is deactivating the Chinese data centers?

Dwarkesh Patel
This is similar to the story you could tell about Big Tech right now. Satya, if he wanted to, could get his engineers to find zero days in Windows and infiltrate the president's computer. Right now, Satya could do that.

Leopold Aschenbrenner
They'd be shut down.

Dwarkesh Patel
What do you mean?

Leopold Aschenbrenner
The government wouldn't let them do that.

Dwarkesh Patel
There's a story where they could pull off a coup.

Leopold Aschenbrenner
They could not pull off a coup.

Dwarkesh Patel
Fine, I agree. What's wrong with a scenario where multiple companies are going for it? The AI is still broadly deployed. Alignment works. The system-level prompt is that it can't help people make bioweapons or something. It's still broadly deployed.

Leopold Aschenbrenner
I expect AIs to be broadly deployed.

Dwarkesh Patel
Even if it's a government project?

Leopold Aschenbrenner
Yeah, I think the Metas of the world open-sourcing their AIs that are two years behind is super valuable. There will be some question of whether the offense-defense balance is fine, and open-sourcing two-year-old AIs is fine. Or there are restrictions on the most extreme dual-use capabilities, like not letting private companies sell crazy weapons.

That's great and will help with diffusion. After the government project, there will be an initial tense period. Hopefully, it stabilizes. Then, like Boeing, they'll go out and do all the flourishing civilian applications like nuclear energy. The civilian applications will have their day.

Dwarkesh Patel
How does that proceed? Because in the other world, there are existing stocks of capital that are worth a lot.

Leopold Aschenbrenner
There will still be Google clusters.

Dwarkesh Patel
So Google, because they got the contract from the government, will control the ASI? But why are they trading with anybody else?

Leopold Aschenbrenner
It'll be the same companies that would be doing it anyway. In this world, they're just contracting with the government or are DPA'd so all their compute goes to the government. In some sense it's very natural.

Dwarkesh Patel
After you get the ASI and we're building the robot armies and fusion reactors…

Leopold Aschenbrenner
Only the government will get to build robot armies.

Dwarkesh Patel
Now I'm worried. Or like the fusion reactors and stuff.

Leopold Aschenbrenner
That's what we do with nukes today.

Dwarkesh Patel
If you already have the robot armies and everything, the existing society doesn't have some leverage where it makes sense for the government to —

Leopold Aschenbrenner
They don't have that today.

Dwarkesh Patel
They do, in the sense that they have a lot of capital that the government wants. There are other things as well. Why was Boeing privatized after WWII?

Leopold Aschenbrenner
The government has the biggest guns. The way we regulate is through institutions, constitutions, legal restraints, courts, etc.

Dwarkesh Patel
Tell me what privatization looks like in the ASI world afterwards.

Leopold Aschenbrenner
Afterwards, it's like the Boeing example.

Dwarkesh Patel
Who gets it?

Leopold Aschenbrenner
Google and Microsoft, the AI labs —

Dwarkesh Patel
Who are they selling it to? They already have robot factories. Why are they selling it to us? They don't need anything from us. This is chump change in the ASI world because we didn't get the ASI broadly deployed throughout this takeoff.

We don't have the robots, the fusion reactors, or the advanced decades of science you're talking about. What are they trading with us for?

Leopold Aschenbrenner
Trading with whom for?

Dwarkesh Patel
For everybody who was not part of the project. They've got technology that's decades ahead.

Leopold Aschenbrenner
That's a whole other issue of how economic distribution works. I don't know. That'll be rough. I'm just saying, I don't see the alternative. The alternative is overturning a 500-year civilizational achievement of Landfrieden. You basically instantly leak the stuff to the CCP.

Either you barely scrape ahead, but you're in a fever struggle, proliferating crazy WMDs. It's enormously dangerous for alignment because you're in a crazy race at the end, and you don't have the ability to take six months to get alignment right. The alternative is not bundling efforts to win the race against authoritarian powers.

I don't like it. I wish we used the ASI to cure diseases and do all the good in the world. But it's my prediction that in the end game, what's at stake is not just cool products but whether liberal democracy survives, whether the CCP survives.

What will the world order for the next century be? When that is at stake, forces will be activated that are way beyond what we're talking about now. In this crazy race at the end, the national security implications will be the most important.

To go back to World War II, nuclear energy had its day, but in the initial period when the technology was first discovered, you had to stabilize the situation. You had to get nukes and do it right. Then the civilian applications had their day.

Dwarkesh Patel
I agree that nuclear energy is a thing that happened later on and is dual-use. But it's something that happened literally a decade after nuclear weapons were developed.

Leopold Aschenbrenner
Right because everything took a long time.

Dwarkesh Patel
Whereas with AI, all the applications are immediately unlocked. This is closer to the analogy people make about AGI. Assume your society had 100 million more John von Neumanns.

If that literally happened, if tomorrow you had 100 million more of them, I don't think the approach would be that we have to worry about some of them converting to ISIS or "what if a bunch are born in China?" I don't think we'd be talking about nationalizing all the John von Neumanns.

I think it'd generally be a good thing. I'd be concerned about one power getting all the John von Neumanns.

Leopold Aschenbrenner
The issue is bottling up, in a short period of time, this enormous unfolding of technological progress, an industrial explosion. We do worry about 100 million John von Neumanns.

Why do we worry about the rise of China? It's one billion people who can do a lot of industry and technology. This is like the rise of China multiplied by 100. It's not just one billion people, but a billion super-intelligent beings. Plus, it comes all in a very short period.

Dwarkesh Patel
Practically, if the goal is to beat China, part of that is protecting ourselves.

Leopold Aschenbrenner
Beating China is just one of the goals. We also want to manage this incredibly crazy, scary period.

Dwarkesh Patel
Right. Part of that is making sure we're not leaking algorithmic secrets to them.

Leopold Aschenbrenner
Building the trillion-dollar cluster.

Dwarkesh Patel
That's right. But isn't your point that Microsoft can issue corporate bonds?

Leopold Aschenbrenner
Microsoft can do hundreds of billions of dollars. The trillion-dollar cluster is closer to a national effort.

Dwarkesh Patel
I thought your earlier point was that American capital markets are deep and good.

Leopold Aschenbrenner
They're pretty good. The trillion-dollar cluster is possible privately, but it's going to be tough.

Dwarkesh Patel
At this point, we have AGI that's rapidly accelerating productivity.

Leopold Aschenbrenner
The trillion-dollar cluster will be planned before the AGI. You get the AGI on the 10 GW cluster. Maybe have one more year of final unhobbling to fully unlock it. Then you have the intelligence explosion.

Meanwhile, the trillion-dollar cluster is almost finished. You run your superintelligence on it. You also have hundreds of millions of GPUs on inference clusters everywhere.

Dwarkesh Patel
In this world, I think private companies have their capital and can raise capital.

Leopold Aschenbrenner
You will need the government to do it fast.

Dwarkesh Patel
We know private companies are on track to do this. In China, if they're unhindered by climate change or whatever —

Leopold Aschenbrenner
That's part of what I'm saying.

Dwarkesh Patel
If it really matters that we beat China…

There will be all sorts of practical difficulties. Will the AI researchers actually join the AI effort? If they do, there will be at least three different teams currently doing pre-training in different companies.

Who decides at some point that you're going to have to YOLO the hyperparameters. Who decides that? Merging extremely complicated research and development processes across very different organizations is somehow supposed to speed up America against the Chinese?

Leopold Aschenbrenner
Brain and DeepMind merged. It was a little messy, but it was fine.

Dwarkesh Patel
It was pretty messy. It was also the same company and much earlier on in the process.

Leopold Aschenbrenner
Pretty similar, right? Different codebases, lots of different infrastructure and teams. It wasn't the smoothest process, but DeepMind is doing very well.

Dwarkesh Patel
You give the example of COVID. In the COVID example, we woke up to it, maybe it was late, but then we deployed all this money. The COVID response from the government was a clusterfuck. I agree that Warp Speed was enabled by the government, but it was literally just giving permission that you can actually —

Leopold Aschenbrenner
It was also making big advance market commitments.

Dwarkesh Patel
I agree. But fundamentally, it was a private sector-led effort. That was the only part of the COVID response that worked.

Leopold Aschenbrenner
The project will look closer to Operation Warp Speed. You'll have all the companies involved in the government project. I'm not convinced that merging is that difficult. You run pre-training on GPUs with one codebase, then do the secondary step on the other codebase with TPUs. It's fine.

On whether people will sign up for it, they wouldn't sign up for it today. It would seem crazy to people.

But this is part of the secrecy thing. People gather at parties and… you know this. I don't think anyone has really gotten up in front of these people and said, "Look, what you're building is the most important thing for the national security of the United States - for the future of the free world and whether we have another century ahead of it. This is really important for your country and democracy. Don't talk about the secrets." It's not just about DeepMind or whatever. It's about these really important things.

We're talking about the Manhattan Project. It was really contentious initially, but at some point it became clear that this was coming. There was an exigency on the military national security front. A lot of people will come around.

On whether it'll be competent, a lot of this stuff is more predictive. This is reasonably likely, and not enough people are thinking about it. A lot of people think about AI lab politics but nobody has a plan for the grand project.

Dwarkesh Patel
Should they be more pessimistic about it? We don't have a plan for it, and we need to act soon because AGI is upon us. The only competent technical institutions capable of making AI right now are private.

Leopold Aschenbrenner
Companies will play that leading role. It'll be a partnership.

We talked about World War II and American unpreparedness. The beginning of World War II was complete shambles. America has a very deep bench of incredibly competent managerial talent. There are a lot of dedicated people. An Operation Warp Speed-like public-private partnership is what I imagine it would look like.

Dwarkesh Patel
Recruiting talent is an interesting question. For the Manhattan Project, you initially had to convince people to beat the Nazis and get on board. Many of them regretted how much they accelerated the bomb. This is generally a thing with war.

Leopold Aschenbrenner
I think they were wrong to regret it.

Dwarkesh Patel
Why?

Leopold Aschenbrenner
What's the reason for regretting it?

Dwarkesh Patel
They way nuclear weapons were developed after the war was explosive because there was a precedent that you can use nuclear weapons. Then because of the race that was set up, you immediately go to the H-bomb.

Leopold Aschenbrenner
This is related to my view on AI and maybe where we disagree. That was inevitable. There was a world war, then a cold war. Of course the military angle would be pursued with ferocious intensity. There's no world in which we all decide not to build nukes. Also, nukes went really well. That could have gone terribly.

It's not physically possible to have something like pocket nukes for everybody, where WMDs proliferated and were fully democratized. The US led on nukes and built a new world order, with a few great powers and a non-proliferation regime for nukes. It was a partnership and a deal: no military application of nuclear technology, but help with civilian technology. They enforced safety norms on the rest of the world. That worked and could have gone much worse.

Not to mention, I say this in the piece but the A-bomb in Hiroshima was just like firebombing. What changed the game was the H-bombs and ICBMs. That's when it went to a whole new level.

Dwarkesh Patel
When you say we will tell people that we need to pursue this project for the free world to survive, it sounds similar to World War II. World War II is a sad story, not only because it happened, but the victory was sad in the sense that Britain went in to protect Poland. At the end, the USSR, which as your family knows is incredibly brutal, ends up occupying half of Europe. The idea of protecting the free world by rushing AI might end up with an American AI Leviathan. We might look back on this with the same twisted irony as Britain going into World War II to protect Poland.

Leopold Aschenbrenner
There will be a lot of unfortunate things that happen. I'm just hoping we make it through. The pitch won't only be about the race. The race will be a backdrop. It's important that democracy shapes this technology. We can't leak this stuff to North Korea.

Safety, including alignment and the creation of new WMDs, is also just important. I'm not convinced there's another path. Say we have a breakneck race internationally, instantly leaking all this stuff, including the weights, with a commercial race with Demis, Dario, and Sam all wanting to be first. It's incredibly rough for safety.

Safety regulation, as people talk about it, is like NIST involving years of bureaucracy and expert consensus.

Dwarkesh Patel
Isn't that what's going to happen with the project as well?

Leopold Aschenbrenner
Alignment during the intelligence explosion is not a years-long bureaucratic process. It's more like a war, with a fog of war. Is it safe to do the next OOM? We're three OOMs into the intelligence explosion, and we don't fully understand what's happening.

Our generalization-scaling curves don’t look great, some automated AI researchers say it's fine, but we don't quite trust them. AIs might start doing problematic things, but we hammer it out, and then it's fine. Should we go ahead? Should we take another six months?

Meanwhile, China might steal the weights or deploy a robot army. It's a crazy situation, relying more on a sane chain of command than a deliberative regulatory scheme. Although I wish we could do that more deliberative regulatory scheme.

This is the thing with private companies too. Private companies claim they'll do safety, but it’s rough in a commercial race, especially for startups. Startups are startups. They aren't fit to handle WMDs.

Dwarkesh Patel
I’m coming closer to your position but…

Let’s talk about the responsible scaling policies. I was told by people advancing this idea — because they know I’m a libertarian-type person and the way they approached me was like this — that the way to think about it was that it's fundamentally a way to protect market-based development of AGI. If you didn’t have this, there would be misuse and lead to nationalization. The RSPs are a way to ensure a market-based order with safeguards to prevent things from going off the rails.

It seems like your story is self-consistent. I know this was never your position, so I’m not looping you into this. But it's almost like a motte-and-bailey argument.

Leopold Aschenbrenner
Here’s what I think about RSP-type stuff and current safety regulations. They’re important for helping us figure out what world we’re in and flashing the warning signs when we’re close.

The story we’ve been telling is what I think the modal version of this decade is. There are many ways it could be wrong. We should talk about the data wall more. There’s a world where this stuff stagnates or we don’t have AGI.

The RSPs preserve optionality. Let’s see how things go, but we need to be prepared if the red lights start flashing. If we get the automated AI researcher, then it’s crunch time.

Dwarkesh Patel
I can be on the same page with that and have a strong prior on pursuing a market-based way. Unless you're right about what the intelligence explosion looks like, don't move yet. But in that world where it really does seem like Alec Radford can be automated, and that's the only bottleneck to getting to ASI…

Okay I think we can leave it at that. I'm somewhat on the way there.

Leopold Aschenbrenner
I hope it goes well. It's going to be very stressful. Right now is the chill time. Enjoy your vacation while it lasts.

Dwarkesh Patel
It's funny to look out over this. This is San Francisco.

Leopold Aschenbrenner
Yeah OpenAI is right there. Anthropic is there. You guys have this enormous power over how it's going to go for the next couple of years, and that power is depreciating.

Dwarkesh Patel
Who's "you guys"?

Leopold Aschenbrenner
People at labs.

It's a crazy world you're talking about. You mention that maybe they'll nationalize too soon. Almost nobody sees what's happening. This is what I find stressful about all this.

Maybe I'm wrong, but if I'm right, we're in this crazy situation where only a few hundred guys are paying attention. It's daunting.

Dwarkesh Patel
I went to Washington a few months ago. I was talking to people doing AI policy stuff there. I asked how likely they think nationalization is. They said it's really hard to nationalize stuff. It's been a long time since it's been done. There are specific procedural constraints on what kinds of things can be nationalized.

Then I asked about ASI. Because of constraints like the Defense Production Act, that won't be nationalized? The Supreme Court would overturn that? They were like, "yeah I guess that would be nationalized."

Leopold Aschenbrenner
That's the short summary of my post or my view on the project.

Dwarkesh Patel
Before we go further on the AI stuff, let's back up.

We began the conversation, and I think people will be confused. You graduated valedictorian of Columbia when you were 19. So, you got to college when you were 15.

You were in Germany then, and you got to college at 15.

Leopold Aschenbrenner
Yeah.

Dwarkesh Patel
How the fuck did that happen?

Leopold Aschenbrenner
I really wanted out of Germany. I went to a German public school. It was not a good environment for me.

Dwarkesh Patel
In what sense? No peers?

Leopold Aschenbrenner
There's also a particular German cultural sense. In the US, there are amazing high schools and an appreciation of excellence. In Germany, there's a tall poppy syndrome. If you're the curious kid in class wanting to learn more, instead of the teacher encouraging you, they resent you and try to crush you.

There are also no elite universities for undergraduates, which is kind of crazy. The meritocracy was crushed in Germany at some point. There's also an incredible sense of complacency across the board. It always puzzles me but even going to a US college was seen as a radical act. It doesn't seem radical to anyone here because it's the obvious thing to do. You can go to Columbia and get a better education.

It's wild to me because this is where stuff is happening and you can get a better education but people in Germany don't do it. I skipped a few grades, and it seemed normal to me at the time to go to college at 15 and come to America. One of my sisters is turning 15 now, and when I look at her, I understand why my mother was worried.

Dwarkesh Patel
So you were presumably the only 15-year-old. Was it normal for you to be a 15-year-old in college? What were the initial years like?

Leopold Aschenbrenner
Again, it felt so normal at the time. Now I understand why my mother was worried. I worked on my parents for a while and eventually persuaded them. It felt very normal at the time.

It was great. I really liked college. It came at the right time for me. I really appreciated the liberal arts education, the core curriculum, and reading core works of political philosophy and literature.

Dwarkesh Patel
You did what? Econ?

Leopold Aschenbrenner
My majors were math, statistics, and economics, but Columbia has a pretty heavy core curriculum and liberal arts education. Honestly, I shouldn't have done all the majors. The best courses were those with amazing professors in some history classes. That's what I would recommend people spend their time on in college.

Dwarkesh Patel
Was there one professor or class that stood out?

Leopold Aschenbrenner
A few. Richard Betts' class on war, peace, and strategy. Adam Tooze was fantastic and has written very riveting books. You should have him on the podcast, by the way.

Dwarkesh Patel
I've tried. I think you tried for me.

Leopold Aschenbrenner
You've got to get him on the pod. It'd be so good.

Dwarkesh Patel
Recently, we were talking to Tyler Cowen. He said when he first encountered you, it was through your paper on economic growth and existential risk. He said, "when I read it, I couldn't believe that a 17-year-old had written it. If this were an MIT dissertation, I'd be impressed." You're a junior and you're writing novel economic papers? Why did you get interested in this, and what was the process to get into that?

Leopold Aschenbrenner
I just get interested in things. It feels natural to me. I get excited about something, read about it, and immerse myself. I can learn and understand information quickly.

Regarding the paper, moments of peak productivity matter more than average productivity, at least for the way at work. Some jobs, like CEO, require consistent productivity. I have periods of a couple months where there's effervescence and other times, I'm computing stuff in the background. Writing the series was similar. You write it and it's really flowing. That's what ends up mattering.

Dwarkesh Patel
Even for CEOs, peak productivity might be very important. One of our friends in a group chat, following Chatham House rules, pointed out how many famous CEOs and founders have been bipolar or manic. The call option on your productivity is the most important thing, and you get it by increasing volatility through being bipolar. That's interesting.

You got interested in economics first. Why economics? You could read about anything. You kind of got a slow start on ML. You wasted all these years on econ. There's an alternative world where you're on the superalignment team at 17 instead of 21 or whatever it was.

Leopold Aschenbrenner
In some sense, I'm still doing economics. I'm looking at straight lines on a graph, log-log plots, figuring out trends, and thinking about feedback loops, equilibrium, and arms control dynamics. It's a way of thinking that I find very useful.

Dario and Ilya seeing scaling early is, in some sense, a very economic way of thinking. It's also related to empirical physics. Many of them are physicists. Economists often can't code well enough, which is their issue, but it's that way of thinking.

I also thought a lot of core ideas in economics were beautiful. In some sense, I feel a little duped because econ academia is kind of decadent now. The paper I wrote is long, 100 pages of math, but the core takeaway can be explained in 30 seconds and it makes sense and you don't really need the math. The best pieces of economics are like that.

You do the work to uncover insights that weren't obvious to you before. Once you've done the work, some sort of mechanism falls out of it that makes a lot of crisp, intuitive sense and explains facts about the world. You can then use it in arguments. Econ 101 is great like this. A lot of econ in the fifties and sixties was like this. Chad Jones' papers are often like this. I really like his papers for this.

Why didn’t I ultimately pursue econ academia? There were several reasons, one of them being Tyler Cowen. He took me aside and said, "I think you’re one of the top young economists I’ve ever met, but you should probably not go to grad school."

Dwarkesh Patel
Oh, interesting. Really? I didn’t realize that.

Leopold Aschenbrenner
Yeah, it was good because he kind of introduced me to the Twitter weirdos. I think the takeaway from that was that I have to move out west one more time.

Dwarkesh Patel
Wait Tyler introduced you to the Twitter weirdos?

Leopold Aschenbrenner
A little bit. Or just kind of the broader culture?

Dwarkesh Patel
A 60-year-old economist introduced you to Twitter?

Leopold Aschenbrenner
Well, I had been in Germany, completely on the periphery, and then moved to a US elite institution. I got a sense of meritocratic elite US society. Basically, there was a directory. To find the true American spirit I had to come out here.

The other reason I didn’t become an economist, or at least pursue econ academia, is that econ academia has become a bit decadent. Maybe it's just that ideas are getting harder to find, or that all the beautiful, simple things have been discovered.

But what are econ papers these days? They’re often 200 pages of empirical analyses on things like how buying 100,000 more textbooks in Wisconsin affects educational outcomes. I'm happy that work happens. It’s important work but it doesn't uncover fundamental insights and mechanisms in society.

Even the theory work often involves really complicated models and the model spits out something like, “Fed does X, then Y happens” and you have no idea why that happened. There’s a gazillion parameters and they’re all calibrated in some way and it’s a computer simulation and you have no idea about the validity. The most important insights are the ones where you have to do a lot of work to get them but then there’s this crisp intuition.

Dwarkesh Patel
The P versus NP of…

Leopold Aschenbrenner
Sure, yeah.

Dwarkesh Patel
That’s really interesting. Going back to your time in college, you say that peak productivity explains this paper and things. But being valedictorian, getting straight A’s, is very much an average productivity phenomenon.

Leopold Aschenbrenner
There’s one award for the highest GPA, which I won, but the valedictorian is selected by the faculty from among those with the highest GPA.

Dwarkesh Patel
So it's not just peak productivity.

Leopold Aschenbrenner
I generally just love this stuff. I was curious, found it really interesting, and enjoyed learning about it. It made sense to me, and it felt very natural.

One of my faults is that I’m not that good at eating glass. Some people are very good at it. The moments of peak productivity come when I’m excited and engaged and love it. If you take the right courses, that’s what you get in college.

Dwarkesh Patel
It’s like Bruce Banner’s quote in The Avengers: "I’m always angry." I’m always excited and curious. That’s why I’m always at peak productivity.

By the way, when you were in college, I was also in college. Despite being a year younger than me, you were ahead of me by at least two years or something. We met around this time through the Tyler Cowen universe. It’s very insane how small the world is. Did I reach out to you? I must have.

Leopold Aschenbrenner
I’m not sure.

Dwarkesh Patel
When I had a couple of videos with a few hundred views.

Leopold Aschenbrenner
It’s a small world. This is the crazy thing about the AI world. It’s the same few people at the parties running the models at DeepMind, OpenAI, and Anthropic. Some of our friends, now

successful in their careers, met many of the people who are now successful in Silicon Valley before their twenties or in their early twenties.

Why is it a small world? There's some amount of agency. I think in a funny way, this is what I took away from my Germany experience. It was crushing. I didn't like it. Skipping grades and moving to the US were unusual moves.

Just trying to do it, and then seeing it work, reinforced the idea that you don't have to conform to the Overton window. You can try to do what seems right to you, even if most people are wrong. That was a valuable and formative early experience.

Dwarkesh Patel
After college, what did you do?

Leopold Aschenbrenner
I did econ research for a bit, at Oxford and other places, and then I worked at Future Fund.

Dwarkesh Patel
Tell me about it.

Leopold Aschenbrenner
Future Fund was a foundation funded by Sam Bankman-Fried but we were our own thing. We were based in the Bay Area. At the time, in early 2022, it was an incredibly exciting opportunity. It was basically a startup foundation, which doesn't come along often. We thought we would be able to give away billions of dollars and remake how philanthropy is done from first principles.

We thought we'd have significant impact, focusing on causes like biosecurity, AI, and finding exceptional talent to work on hard problems. A lot of the work we did was exciting. Academics, who would usually take six months, would send us emails saying, "This is great. This is so quick and straightforward." I often find that with a little encouragement and empowerment, by removing excuses and making the process easy, you can get people to do great things.

Dwarkesh Patel
For context, not only were you guys planning on deploying billions of dollars, but it was a team of four people. So you, at 18, were on a team of four people that was in charge of deploying billions of dollars.

Leopold Aschenbrenner
That was sort of the heyday. Then in November 2022, it was revealed that Sam was a giant fraud, and from one day to the next, the whole thing collapsed. It was really tough. It was

devastating for the people who had their money in FTX. Closer to home, we wanted to help all the grantees do amazing projects but they ended up suddenly saddled with a giant problem.

Personally, it was difficult because it was a startup. I had worked 70-hour weeks every week for almost a year to build it up. We were a tiny team, and then from one day to the next, it was all gone and associated with a giant fraud. That was incredibly tough.

Dwarkesh Patel
Were there any early signs about SBF?

Leopold Aschenbrenner
Obviously, I didn't know he was a fraud. If I had, I would have never worked there. We were a separate entity and didn't work with the business. I do think there are some takeaways for me.

I, and people in general, had this tendency to give successful CEOs a pass on their behavior because they're successful. You think that's just a successful CEO thing. I didn't know Sam Bankman-Fried was a fraud.

I knew he was extremely risk-taking, narcissistic, and didn't tolerate disagreement well. By the end, he and I didn't get along because I pointed out that some biosecurity grants weren't cost effective but he liked them because they were cool and flashy. He was unhappy about that.

So I knew his character. I realized that it's really worth paying attention to people's characters, including people you work for and successful CEOs. That can save you a lot of pain down the line.

Dwarkesh Patel
After FTX imploded and you were out, you went to OpenAI. The superalignment team had just started. You were part of the initial team.

What was the original idea? What compelled you to join?

Leopold Aschenbrenner
The alignment teams at OpenAI and other labs had done basic research and developed RLHF. reinforcement learning from human feedback. That ended up being a really successful technique for controlling current AI models.

Our task was to find the successor to RLHF. The reason we need that is that RLHF probably won't scale to superhuman systems. RLHF relies on human raters giving feedback, but superintelligent models will produce complex outputs beyond human comprehension. It'll be like a million lines of complex code and you won't know at all what's going on anymore.

How do you steer and control these systems? How do you add side constraints? I joined because I thought this was an important and solvable problem. I still do and even more so. I think there's a lot of promising ML research on aligning superhuman systems, which we can discuss more later.

Dwarkesh Patel
It was so solvable, you solved it in a year. It's all over now.

Leopold Aschenbrenner
OpenAI wanted to do a really ambitious effort on alignment. Ilya was backing it. I liked a lot of the people there. I was really excited. There are always people making hay about alignment. I appreciate people highlighting the importance of the problem and I was just really into trying to solve it. I wanted to do the ambitious effort, like an Operation Warp Speed for solving alignment. It seemed like an amazing opportunity to do it.

Dwarkesh Patel
Now the team basically doesn't exist. The heads of it, Jan and Ilya, have left. That's been the news of last week. What happened? Why did the team break down?

Leopold Aschenbrenner
OpenAI decided to take things in a different direction.

Dwarkesh Patel
Meaning what? That superalignment isn't the best way to frame it?

Leopold Aschenbrenner
No, obviously after the November board events there were personnel changes. Ilya leaving was incredibly tragic for OpenAI. There was some reprioritization. There's been reporting on the superalignment compute commitment, the 20% compute commitment, which was how a lot of people were recruited. There was a decision to not keep that commitment and go in a different direction.

Dwarkesh Patel
Now Jan and Ilya have left, and the team itself has dissolved. You were the first person who left or was forced to leave. The Information reported that you were fired for leaking. What happened? Is this accurate?

Leopold Aschenbrenner
Why don't I tell you what they claim I leaked, and you can tell me what you think. OpenAI claimed to employees that I was fired for leaking. I and others have pushed them to say what the leak was. Here's their response in full: Sometime last year, I had written a brainstorming document on preparedness, safety, and security measures needed in the future on the path to AGI. I shared that with three external researchers for feedback. That's the leak.

For context, it was totally normal at OpenAI at the time to share safety ideas with external researchers for feedback. It happened all the time. The doc had my ideas. Before I shared it, I reviewed it for anything sensitive. The internal version had a reference to a future cluster, which I redacted for the external copy. There was a link to some internal slides, but that was a dead link for the external people. The slides weren't shared with them.

When I pressed them to specify what confidential information was in this document. They came back with a line about planning for AGI by 2027-2028 and not setting timelines for preparedness.

I wrote this doc a couple of months after the superalignment announcement. We had put out a four-year planning horizon. I didn't think that planning horizon was sensitive. It's the sort of thing Sam says publicly all the time. I think Jan mentioned it on a podcast a couple of weeks ago. So, that's it.

Dwarkesh Patel
That's it? That sounds pretty thin if the cause was leaking. Was there anything else to it?

Leopold Aschenbrenner
That was the leaking claim. Let me explain more about what happened during the firing. Last year, I wrote an internal memo about OpenAI's security, which I thought was egregiously insufficient to protect against the theft of model weights or key algorithmic secrets from foreign actors. I shared this memo with a few colleagues and a couple of members of leadership, who mostly said it was helpful.

A few weeks later, a major security incident occurred1. That prompted me to share the memo with a couple of board members. Days later, it was made very clear to me that leadership was very unhappy I had shared this memo with the board. Apparently, the board hassled leadership about security.

I got an official HR warning for sharing the memo with the board. The HR person told me it was racist to worry about CCP espionage and that it was unconstructive. I probably wasn't at my most diplomatic and could have been more politically savvy. I thought it was a really important issue. The security incident made me very worried.

The reason I bring this up is that when I was fired, it was very made explicit that the security memo was a major reason for my being fired. They said, "the reason this is a firing and not a warning is because of the security memo."

Dwarkesh Patel
You sharing it with the board?

Leopold Aschenbrenner
The warning I'd gotten for the security memo.

What might also be helpful context is the kinds of questions they asked me when they fired me. A bit over a month ago, I was pulled aside for a chat with a lawyer that quickly turned adversarial. The questions were about my views on AI progress, on AGI, the appropriate level of security for AGI, whether the government should be involved in AGI, whether I and the superalignment team were loyal to the company, and what I was up to during the OpenAI board events. They then talked to a couple of my colleagues and came back and told me I was fired. They'd gone through all of my digital artifacts from my time at OpenAI, and that's when they found the leak.

The main claim they made was this leaking allegation. That's what they told employees. The security memo was another thing. There were a couple of other allegations they threw in. One thing they said was that I was unforthcoming during the investigation because I didn't initially remember who I had shared the preparedness brainstorming document with, only that I had talked to some external researchers about these ideas.

The document was over six months old, I'd spent a day on it. It was a Google Doc I shared with my OpenAI email. It wasn't a screenshot or anything I was trying to hide. It simply didn't stick because it was such a non-issue. They also claimed I was engaging on policy in a way they didn't like. They cited there that I had spoken to a couple of external researchers, including someone at a think tank, about my view that AGI would become a government project, as we just discussed.

In fact, I was speaking with lots of people in the field about that view at the time. I thought it was a really important thing to think about. So they found a DM I had written to a friendly colleague, five or six months earlier, and they cited that too. I had thought it was well within OpenAI norms to discuss high-level issues about the future of AGI with external people in the field.

That's what they allege happened. I've spoken to a few dozen former colleagues about this since. The universal reaction has been, "that's insane." I was surprised as well. I had been promoted just a few months before. Ilya's comment for the promotion case at the time was something like, "Leopold's amazing. We're lucky to have him."

The thing I understand, and in some sense it's reasonable, is that I ruffled some feathers and was probably annoying at times with the security stuff. I repeatedly raised that, maybe not always in the most diplomatic way. I didn't sign the employee letter during the board events, despite pressure to do so.

Dwarkesh Patel
You were one of like eight people or something?

Leopold Aschenbrenner
Not that many people. I think the two most senior people who didn't sign were Andrej and Jan, who have both since left.

On the letter, by Monday morning when it was circulating, I thought it was probably appropriate for the board to resign because they had lost too much credibility and trust with the employees.

But I thought the letter had issues. It didn't call for an independent board, which is a basic of corporate governance. In other discussions, I pressed leadership for OpenAI to abide by its public commitments. I raised tough questions about whether it was consistent with the OpenAI mission and the national interest to partner with authoritarian dictatorships to build the core infrastructure for AGI.

It's a free country. That's what I love about it. We talked about it. They have no obligation to keep me on staff. It would have been reasonable for them to come to me and say, "we're taking the company in a different direction. We disagree with your point of view. We don't trust you to toe the company line anymore. Thank you so much for your work at OpenAI, but it's time to part ways."

That would have made sense. We had started diverging on important issues. I came in very excited and aligned with OpenAI, but that changed over time. That would have been a very amicable way to part ways. It's a shame how it went down.

All that being said, I really want to emphasize that there are a lot of incredible people at OpenAI, and it was an incredible privilege to work with them. Overall, I'm extremely grateful for my time there.

Dwarkesh Patel
Now there's been reporting about an NDA that former employees have to sign to access their vested equity. Did you sign such an NDA?

Leopold Aschenbrenner
No. My situation was a little different because I was right before my cliff. They still offered me the equity, but I didn't want to sign. Freedom is priceless

Dwarkesh Patel
How much was the equity?

Leopold Aschenbrenner
Close to a million dollars.

Dwarkesh Patel
So it was definitely something you and others were aware of. OpenAI explicitly offered you a choice. Presumably, the person on OpenAI staff knew they were offering equity but required signing an NDA that prevents making statements about AGI and OpenAI, like the ones you're making on this podcast.

Leopold Aschenbrenner
I don't know the whole situation. I certainly think conditioning vested equity on signing an NDA is pretty rough. It might be different if it's a severance agreement.

Dwarkesh Patel
Right, but an OpenAI employee who had signed it presumably couldn't give the podcast you're giving today.

Leopold Aschenbrenner
Quite possibly not. I don't know.

Dwarkesh Patel
The board thing is really tough. Analyzing the situation here, if you were trying to defend them, you might say, "well, listen you were just going outside the regular chain of command." There might be a point there.

Although the idea that HR thinks you're supposed to have an adversarial relationship with the board is odd. You're giving the board relevant information about whether OpenAI is fulfilling its mission and how it can improve. That seems important since the board is supposed to ensure OpenAI follows its mission. Them treating that as part of the leak, as if the board were an external actor…

Leopold Aschenbrenner
To be clear, the leak allegation was just about that document I shared for feedback. This is a separate issue they cited. They said I wouldn't have been fired if not for the security memo.

Dwarkesh Patel
They said you wouldn't have been fired for it.

Leopold Aschenbrenner
They said the reason this is a firing and not a warning is because of the warning I had gotten for the security memo.

Dwarkesh Patel
Before you left, the incidents with the board happened. Sam was fired and then rehired as CEO, and now he's on the board. Ilya and Jan, who were the heads of the superalignment team, have left. Ilya, was a co-founder of OpenAI and the most significant member of OpenAI from a research perspective. There has been a lot of personnel drama over the last few months regarding superalignment and just generally with the OpenAI personnel drama. What's going on?

Leopold Aschenbrenner
There's a lot of drama. Why is there so much drama?

There would be much less drama if all OpenAI claimed to be was building ChatGPT or business software. A lot of the drama comes from OpenAI really believing they're building AGI. That isn't just a marketing claim. There's a report that Sam is raising $7 trillion for chips. That only makes sense if you really believe in AGI.

What gets people is the cognitive dissonance between believing in AGI and not taking some of the other implications seriously. This technology will be incredibly powerful, both for good and bad. That implicates national security issues. Are you protecting the secrets from the CCP? Does America control the core AGI infrastructure or does a Middle Eastern dictator control it?

The thing that really gets people is the tendency to make commitments and say they take these issues seriously, but then frequently not follow. For instance, as mentioned, there was a commitment around superalignment compute, dedicating 20% of compute for long-term safety research.

You and I could have a totally reasonable debate about the appropriate level of compute for superalignment. That's not really the issue. The issue is that the commitment was made and it was used to recruit people. It was very public.

It was made because there was a recognition that there would always be something more urgent than long-term safety research, like a new product. In the end, they just didn't keep the commitment. There was always something more urgent than long-term safety research.

Another example is when I raised security issues. They would tell me security is our number one priority. Invariably, when it came time to invest serious resources or make trade-offs to take basic measures, security was not prioritized. The cognitive dissonance and unreliability cause a lot of the drama.

Dwarkesh Patel
Let's zoom out and talk about a big part of the story. A big motivation for the way we must proceed with regards to geopolitics is that once you have AGI, you soon proceed to ASI, or superintelligence. You have these AGIs functioning as researchers into further AI progress and within a matter of years, maybe less, you reach superintelligence. From there, according to your story, you do all this research and development into robotics, pocket nukes, and other crazy shit.

I'm skeptical of this story for many reasons. At a high level, it's not clear to me that this input-output model of research is how things actually happen in research. We can look at the economy as a whole. Patrick Collison and others have pointed out that, compared to 100 years ago, we have 100x more researchers in the world. Yet progress isn't happening 100 times faster. It's clearly not as simple as pumping in more researchers to get higher research output. I don't see why it would be different for AI researchers.

Leopold Aschenbrenner
This is getting into good stuff. This is the classic disagreement I have with Patrick and others. Obviously, inputs matter. The United States produces a lot more scientific and technological progress than Liechtenstein or Switzerland.

Say you made Patrick Collison dictator of Liechtenstein or Switzerland and he implemented his utopia of ideal institutions. Keep the talent pool fixed. He's not able to do some crazy high-skilled immigration thing or genetic breeding scheme. You keep the talent pool fixed with amazing institutions. Even then, even if Patrick Collison were the dictator, Switzerland still wouldn't be able to outcompete the United States in scientific and technological progress. Magnitudes matter.

Dwarkesh Patel
I'm not sure I agree with this. There are many examples in history where small groups of people, Bell Labs or Skunk Works, have made significant progress. OpenAI has a couple hundred researchers.

Leopold Aschenbrenner
Highly selected though.

Dwarkesh Patel
That's why Patrick Collison as a dictator would do a good job of this.

Leopold Aschenbrenner
Well, yes, if he can highly select all the best AI researchers in the world, he might only need a few hundred. But that's the talent pool. You have the 300 best AI researchers in the world.

Dwarkesh Patel
But from 100 years ago to now, the population has increased massively. You would expect the density of talent to have increased, considering that things like malnutrition and poverty which affected past talent are no longer as debilitating to the same level.

Leopold Aschenbrenner
I don't know if it's 100x. It's probably at least 10x. Some people think ideas haven't gotten much harder to find, so why would we need this 10x increase in research effort? To me, this is a very natural story. Why is it natural? It's a straight line on a log-log plot. It's a deep learning researcher's dream.

What is this log-log plot? On the x-axis you have log cumulative research effort. On the y-axis you have log GDP, OOMs of algorithmic progress, log transistors per square inch, log price for a gigawatt of solar energy. It's extremely natural for that to be a straight line. It's classic. Initially, things are easy, but you need logarithmic increments of cumulative research effort to find the next big thing. This is a natural story.

One objection people make is, "isn't it suspicious that we increased research effort 10x and ideas also got 10x harder to find, perfectly equilibrating?" I say it's just equilibrium—it's in a endogenous equilibrium. Isn't it a coincidence that supply equals demand and the market clears? It's the same here. The difficulty of finding new ideas depends on how much progress has been made.

The overall growth rate is a function of how much ideas have gotten harder to find in ratio to how much research effort has increased. This story is fairly natural, and you see it not just economy-wide but also in the experience curve for various technologies.

It's plausible that institutions have worsened by some factor. Obviously, there's some sort of exponent of diminishing returns on adding more people. Serial time is better than just parallelizing. Still, clearly inputs clearly matter.

Dwarkesh Patel
I agree, but if the coefficient, of how fast they diminish as you grow the input, is high enough, then in the abstract the fact that inputs matter isn't that relevant.

We're talking at a very high level, but let's take it down to the concrete. OpenAI has a staff of at most a few hundred directly involved in algorithmic progress for future models. Let's say you could really arbitrarily scale this number for faster algorithmic progress and better AI.

It’s not clear why OpenAI doesn’t just go hire every person with a 150 IQ, of which there are hundreds of thousands in the world.

My story is that there are transaction costs to managing all these people. They don’t just go away if you have a bunch of AIs. These tasks aren’t easy to parallelize. I’m not sure how you would explain the fact that OpenAI doesn’t go on a recruiting binge of every genius in the world?

Leopold Aschenbrenner
Let’s talk about the OpenAI example and the automated AI researchers. Look at the inflation of AI researcher salaries over the last year. It’s gone up by 4x or 5x. They’re clearly trying to recruit the best AI researchers in the world and they do find them. My response would be that almost all of these 150 IQ people wouldn’t just be good AI researchers if you hired them tomorrow. They wouldn’t be Alec Radford.

Dwarkesh Patel
They’re willing to make investments that take years to pan out. The data centers they’re buying now will come online in 2026. Some of them won’t work out, some won’t have traits we like. But why wouldn’t they make the investment to turn these 150 IQ people into amazing AI researchers by 2026?

Leopold Aschenbrenner
Sometimes this does happen. Smart physicists have been really good at AI research, like all the Anthropic co-founders.

Dwarkesh Patel
But for example, Dario said on the podcast that they have a careful policy of being extremely selective and not hiring arbitrarily.

Leopold Aschenbrenner
Training is not as easily scalable. Training is really hard. If you hired 100,000 people, it would be really hard to train them all. You wouldn’t be doing any AI research. There are huge costs to bringing on new people and training them.

This is very different with AIs. It’s important to talk about the advantages AIs will have. What does it take to be an Alec Radford? You need to be a really good engineer. AIs will be amazing engineers and coders. You can train them to do that. They also need to have good research intuitions and a really good understanding of deep learning.

Alec Radford, or people like him, has acquired this over years of being deeply immersed in deep learning and having tried lots of things himself and failed. AIs will be able to read every research paper ever written, learn from every experiment ever run at the lab, and gain intuition from all of this. They'll be able to learn in parallel from each other's experiments and experiences.

There's also a cultural acclimation aspect. If you hire someone new, there's politicking, and maybe they don't fit in well. With AIs, you just make replicas. There's a motivation aspect as well. If I could duplicate Alec Radford, and before I run every experiment, have him spend a decade's worth of human time double-checking code and thinking really carefully about it, he wouldn't care and he wouldn't be motivated. With AIs, you can have 100 million of them focused on making sure the code is correct with no bugs.

The idea of 100 million human-equivalent AI researchers is just a way to visualize it. You might not have literally 100 million copies. There's trade offs you can make between serial speed and parallel. You might run them at 10x or 100x serial speed, resulting in fewer tokens overall because of inherent trade-offs. You might have 100,000 AIs running at 100x human speed. They can coordinate by sharing latent space and attending to each other's context. There's a huge range of possibilities for what you can do.

Another illustration is that by 2027 or 2028, with automated AI researchers, you'll be able to generate an entire Internet's worth of tokens every day. It's clearly a huge amount of intellectual work that you can do.

Dwarkesh Patel
Here's an analogy. Today, we generate more patents in a year than during the actual physics revolution in the early 20th century. Are we making more physics progress in a year today than we did in half a century back then? Generating all these tokens doesn't necessarily equate to generating as much codified knowledge in the initial creation of the Internet.

Leopold Aschenbrenner
Internet tokens are usually final output. We talked about the unhobbling. I think of a GPN token as one token of my internal monologue. That's how I do this math on human equivalents. It's like 100 tokens a minute and then humans working for X hours. What is the equivalent there?

Dwarkesh Patel
This goes back to something from earlier. Why haven't we seen huge revenues from AI yet? People often ask this question. If you took GPT-4 back ten years, people would think it would automate half the jobs. There's a modus ponens, modus tollens here. Part of the explanation is that we're on the verge and we just need to do these unhobblings. Part of that is probably true. But there is another lesson to learn there. Just looking at a set of abilities at face value, there are likely more hobblings behind the scenes. The same will be true of AGIs running as AI researchers.

Leopold Aschenbrenner
I basically agree with a lot of what you said. My story here is that there's going to be a long tail. Maybe by 2026 or 202, you'll have the proto-automated engineer that's really good at engineering. It doesn't yet have the research intuition. You don't quite know how to put them to work.

Even so, the underlying pace of AI progress is already so fast. In just three years, we've gone from AI not being able to do any kind of math at all to now crushing these math competitions. So, you might have the initial automated research engineer by 2026 or 2027, which speeds you up by 2x. You go through a lot more progress in that year. By the end of the year, you've figured out the remaining unhobblings and you've got a smarter model.

Maybe it's two years but then maybe that model can automate 100% of the research. They don't need to be doing everything. They don't need to make coffee or deal with tacit knowledge in other fields. AI researchers at AI labs really know the job of an AI researcher. There are lots of clear metrics. It's all virtual. There's code. There are things you can develop and train for.

Dwarkesh Patel
Another thing is how do you actually manage a million AI researchers? Humans' comparative ability, and we've been especially trained for it, is to work in teams. We've been learning for thousands of years about how we work together in groups. Despite this, management is a clusterfuck. Most companies are poorly managed. It's really hard to do this stuff.

For AIs, we talk about AGI, but for it will be some bespoke set of abilities some of which will be higher than human level and some at human level. It will be some bundle and you'll need to figure out how to put these bundles together with their human overseers and equipment. I'm just very skeptical of the idea that as soon as you get the bundle, you can just shove millions of them together and manage them.

Any other technological revolution in history has been much more piecemeal than you'd expect on paper. What is the industrial revolution? We dug up coal to power steam engines, used steam engines to run railroads, which helped us get more coal. There's sort of a Factorio story you can tell where in six hours you can be pumping out thousands of times more coal. In real life, it takes centuries.

For example, with electrification, there's a famous study showing how it took decades after electricity to switch from the pulley and water wheel-based system for steam engines to one that works with more spread-out electrical motors. This will be the same kind of thing. It might take decades to actually get millions of AI researchers to work together effectively.

Leopold Aschenbrenner
This is great. A few responses to that. I totally agree with the real-world bottlenecks idea. It's easy to underrate these constraints. Basically, we're automating labor and exploiting technology, but there are still many other bottlenecks in the world.

That's why the story starts narrowly where there aren't these bottlenecks and then expands to broader areas over time. This is part of why I think initially it's an AI research explosion. AI research doesn't run into these real-world bottlenecks. It doesn't require plowing a field or digging up coal. It's just doing AI research.

Dwarkesh Patel
I love how in your model, AI research isn't complicated. It's like flipping a burger. It's just AI research.

Leopold Aschenbrenner
People make these arguments like, "AGI won't do anything because it can't flip a burger." Yeah it won't be able to flip a burger, but it'll be able to do algorithmic progress. Once it achieves that, it can figure out how to create a robot that flips burgers. The quantities we're talking about are a lower bound. We can definitely run 100 million of these.

One of the first things we'll figure out is how to translate quantity into quality. Even at the baseline rate of progress, you're quickly getting smarter and smarter systems. It took four years to go from preschooler to high schooler. Pretty quickly, there are probably some simple algorithmic changes you find if you have a hundred Alec Radfors instead of one. You don't even need a hundred million. We'll soon have systems that are even smarter and capable of creative, complicated behavior we don't understand.

Maybe there's some way to use all this test time compute in a more unified way than all these parallel copies. They won't just be quantitatively superhuman. They'll pretty quickly become qualitatively superhuman. It's like a high school student trying to understand standard physics versus a super-smart professor who gets quantum physics. You quickly enter that regime just given the underlying pace of AI progress but even more quickly with the accelerated force of automated AI research.

Dwarkesh Patel
I agree that over time you would get there. I'm not denying that ASI is possible. I'm just questioning how this happens in a year.

Leopold Aschenbrenner
The story is a bit more continuous. By 2025 or 2026, you'll already have models as good as a college graduate. I don't know where all the unhobbling is going to be but even it's possible that you have a proto-automated engineer.

There's a bit of an AGI smear that there are unhobblings missing. There's ways of connecting them that are missing. There's some level of intelligence you're missing. At some point you are going to get this thing that is 100% automated Alec Radford and once you have that, things really take off.

Dwarkesh Patel
Let's go back to the unhobbling.
We're going to get a bunch of models by the end of the year. Suppose we didn't get some capacity by the end of the year. Is there some such capacity which us lacking would suggest that AI progress will take longer than you are projecting?

Leopold Aschenbrenner
There are two key things: the unhobbling and the data wall. Let's talk about the data wall for a moment. Even though we're seeing crazy AI progress, the data wall is actually underrated. There's a real scenario where we stagnate because we've been riding this tailwind of easily bootstrapping unsupervised learning.

It learns these amazing world models. You just buy more compute, make simple efficiency changes, and get big gains. All of the big gains in efficiency have been pretty dumb things. You add a normalization layer. You fix scaling laws. These have already been huge things, let alone obvious ways in which these models aren't good yet.

The data wall is a big deal. For instance, Common Crawl online is about 30 trillion tokens. Llama-3 was trained on 15 trillion tokens. We're already using all of the data. You can get somewhat further by repeating data, but an academic paper by Boaz Barak that does scaling laws for this. It says that after about 16 repetitions, the returns basically go to zero.

Llama-3 is already at the limit of data. Maybe we can get 10x more by repeating data. At most that's a 100x model than GPT-4, a 100x effective compute from GPT-4. That's not that much. If you do half an OOM of compute and half an OOM of algorithmic progress each year, that's like two years from GPT-4. GPT-4 finished pre-training in 2022, so it's 2024. We won't quite know by the end of the year but by 2025 and 2026 we'll get a sense of if we're cracking the data wall.

Dwarkesh Patel
Suppose we had three OOMs less data in Common Crawl on the Internet than we happen to have now. For decades, with the Internet and other things, the stock of data humanity has has been rapidly increasing. Is it your view that, for contingent reasons, we just happen to have enough data to train models just powerful enough, like GPT-4.5, to kick off the self-play RL loop?

Or is it just that if it had been 3 OOMs higher, then progress would have been slightly faster? In that world, we would have been looking back, thinking it would have been hard to kick off the RL explosion with just GPT-4.5. We would have figured it out eventually.

In this world, we would have gotten to GPT-3 and then had to kick off some sort of RL explosion. We would have still figured it out. Did we just luck out on the amount of data we happen to have in the world?

Leopold Aschenbrenner
3 OOMs less data is pretty rough. That would mean 6 OOMs less compute model and Chinchilla scaling laws. That’s basically capping out at something barely better than GPT-2. That would be really rough.

You make an interesting point about contingency. If we consider the human trajectory analogy, a preschooler model can't learn from itself. An elementary school model can't learn from itself. Maybe GPT-4 is like a smart high schooler that can start learning from itself. Ideally, you want a somewhat better model that can truly learn by itself. 1 OOM less data would make me more iffy, but it might still be doable. It would feel chiller if we had 1 or two OOMs of more data.

Dwarkesh Patel
It would be an interesting exercise to get probability distributions of AGI contingent across OOMs of data.

Leopold Aschenbrenner
Yeah, I agree.

Dwarkesh Patel
The thing that makes me skeptical of this story is that it totally makes sense why pre-training works so well. With these other things, there are stories of why they ought to work in principle.. Humans can learn this way and so on. Maybe they're true.

I worry that a lot of this case is based on first principles evaluation of how learning happens. Maybe fundamentally, we don't understand how humans learn. Maybe there's some key thing we're missing. On sample efficiency, you say the fact that these things are way less sample efficient than humans in learning suggests there's a lot of room for improvement. Another perspective is that we are just on the wrong path altogether. That’s why they’re so sample inefficient when it comes to pre-training.

There are a lot of first principles arguments stacked on top of each other where you get these unhobblings, then you get to AGI. Then because of these reasons why you can stack all these things on top of each other and you get to ASI. I'm worried that there are too many steps of this sort of first principles thinking.

Leopold Aschenbrenner
We'll see. On sample efficiency, it’s sort of first principles but there's this clear missing middle. People hadn't been trying. Now people are really trying. Again, often in deep learning something like the obvious thing works and there are a lot of details to get right. It might take some time, but now people are really trying. We will get a lot of signal in the next couple of years on unhobbling.

What is the signal on unhobbling that would be interesting? The question is basically, are you making progress on test time compute? Is this thing able to think longer horizon than just a couple hundred tokens? That was unlocked by chain-of-thought.

Dwarkesh Patel
On that point in particular, many people who have longer timelines have come on the podcast and made the point that the way to train this long horizon RL, it's not…

Earlier we were talking about how they can think for five minutes, but not for longer. It's not because they can't physically output an hour's worth of tokens.

Leopold Aschenbrenner
Even Gemini has a million in context, and the million of context is actually great for consumption. It solves one important hobbling, which is the onboarding problem. A new coworker in your first five minutes, like a new smart high school intern, is not useful at all.

A month in, they’re much more useful because they've looked at the monorepo, understand how the code works, and they've read your internal docs. Being able to put that in context solves this onboarding problem. They're not good at the production of a million tokens yet.

Dwarkesh Patel
On the production of a million tokens, there's no public evidence that there's some easy loss function where you can…

Leopold Aschenbrenner
GPT-4 has gotten a lot better since launch. The GPT-4 gains since launch are a huge indicator.

You talked about this with John Schulman on the podcast. John said this was mostly post-training gains. If you look at the LMSys scores, it's like 100 Elo or something. It's a bigger gap than between Claude 3 Opus and Claude 3 Haiku. The price difference between those is 60x.

Dwarkesh Patel
But it’s not more agentic. It's better in the same chat.

Leopold Aschenbrenner
It’s much better about math. It went from 40% to 70%. That indicates that clearly there's stuff to be done on hobbling. The interesting question is, this time a year from now, is there a model that is able to think for a few thousand tokens coherently, cohesively, identically? Again, I'd probably feel better if we had 1–2 OOMs more data because the scaling just gives you this tailwind.

With tools, when you talk to people who try to make things work with tools, GPT-4 is really when tools start to work. You can kind of make them work with GPT-3.5, but it's just really tough. Having GPT-4, you can help it learn tools in a much easier way. So it’d be great to have just a bit more tailwind from scaling. I don't know if it'll work, but it's a key question.

Dwarkesh Patel
It's a good place to sort of close that part where we know what the crux is and what evidence of that would look like.

Let’s talk about AGI to superintelligence. Maybe it's the case that the gains are really easy right now and you can just sort of let loose. Give Alec Radford a compute budget and he’ll comes out the other end with something that is an additive change as part of the code.

How many other domains in the world are like this, where you think you could get the equivalent of in one year? You just throw enough intelligence across multiple instances and you come out the other end with something that is remarkably decades, centuries ahead? You start off with no flight, and then you have the Wright brothers. You have a million instances of GPT-6, and you come out the other end with Starlink? Is that your model of how things work?

Leopold Aschenbrenner
You're exaggerating the timelines a little bit, but I think a decade's worth of progress in a year or something is a reasonable prompt. This is where the automated AI researcher comes in. It gives you this enormous tailwind on all the other stuff.

You automate AI research with your automated Alec Radfords. You come out the other end. You've done another five OOMs. You have a thing that is vastly smarter. Not only is it vastly smarter, you've been able to make it good at everything else. You're solving robotics.

The robots are important because for a lot of other things, you do actually need to try things in the physical world. Maybe you can do a lot in simulation. Those are the really quick worlds. I don't know if you saw the last Nvidia GTC and it was all about the digital twins having all your manufacturing processes in simulation. Again, if you have these superintelligent cognitive workers, can they just make simulations of everything, off the float style, and make a lot of progress there?

You're just going to get the robots. I agree there are a lot of real-world bottlenecks. It's quite possible that we're going to have crazy drone swarms, but also lawyers and doctors still need to be humans because of regulation. You kind of start narrowly, you broaden, and there are worlds in which you let them loose. Again, because of these competitive pressures, we will have to let them loose to some degree on various national security applications. Rapid progress is quite possible.

In the explosion after, there are basically two components. The A in the production function, the growth of technology, has massively accelerated. Now you have a billion superintelligent scientists and engineers and technicians, superbly competent at everything.

You also just automated labor. Even without the whole technological explosion thing, you have this industrial explosion, at least if you let them loose, because you can cover Nevada and you start with one robot factory producing more robots. It's this cumulative process because you've taken labor out of the equation.

Dwarkesh Patel
That's super interesting.

Although when you increase the K or the L without increasing the A, you can look at examples like the Soviet Union or China. They rapidly increased inputs, which did have a geopolitically game-changing effect. It is remarkable to see the transformation of cities like Shanghai over just decades.

Leopold Aschenbrenner
They throw out these crazy cities in like a decade.

Dwarkesh Patel
People talk about 30% growth rates from AI. The closest thing—

Leopold Aschenbrenner
Look at the Asian Tigers at 10%. It's totally possible.

Dwarkesh Patel
But without productivity gains, it's not like the Industrial Revolution. From a perspective of outside the system, your goods become much cheaper and you can manufacture more things. But it's not a sign that the next century is rapidly approaching.

Leopold Aschenbrenner
Both are important. The other thing I'll say is that with all of this stuff, the magnitudes are really, really important. We talked about a 10x in research effort, or maybe 10-30x over a decade. Even without any kind of self-improvement type loops — even in the sort of GPT-4 to AGI story — we're talking about an OOM of effective compute increase a year.

It's half an OOM of compute, half an OOM of algorithmic progress that sort of translates into effective compute. You're basically doing 10x a year on your labor force. It's a radically different world if you're doing a 10x or 30x in a century versus a 10x a year on your labor force. The magnitudes really matter.

It also really matters in the intelligence explosion scenario, just the automated AI research part. One story you could tell there is that ideas get harder to find. Algorithmic progress is going to get harder. Right now, you have the easy wins, but in like four or five years, there will be fewer easy wins. So the sort of automated AI researchers are going to be necessary to just keep it going, because it's gotten harder. That's sort of a really weird knife-edge assumption economics.

Dwarkesh Patel
Isn't that the equilibrium story you were just telling about why the economy as a whole has 2% economic growth? You just proceed on the equilibrium. I guess you're saying by the time—

Leopold Aschenbrenner
The result of the equilibrium here is that it's way faster. Alt depends on the sort of exponents. Suppose you need to 10x the effective research effort in AI research in the last four or five years to keep the pace of progress. We're not just getting a 10x, you're getting 1,000,000x or 100,000x. The magnitudes really matter.

One way to think about this is that you have two exponentials. You have your normal economy that's growing at 2% a year, and you have your AI economy growing at 10x a year. It's starting out really small. It's way faster and it's going to overtake eventually. You can just do the simple revenue extrapolation if you think your AI economy has some growth rate. It's a very simplistic way, but there's this 10x a year process.

You're going to transition the whole economy, as it broadens, from the 2% a year to the much faster growing process. That's very consistent with historical stories. There's this long-run hyperbolic trend. It manifested in the change in growth mode in the Industrial Revolution, but there's just this long-run hyperbolic trend. Now you have another change in growth mode.

Dwarkesh Patel
That was one of the questions I asked Tyler when I had him on the podcast. The fact that, after 1776, you went from a regime of negligible economic growth to 2% is really interesting. From the perspective of somebody in the Middle Ages or before, 2% is the equivalent of like 10%. I guess you're projecting even higher for the AI economy.

Leopold Aschenbrenner
It depends. Again, with all this stuff I have a lot of uncertainty. A lot of the time I'm trying to tell the modal story because it's important to be concrete and visceral about it.

I have a lot of uncertainty over how the 2030s play out. The thing I know is that it's going to be fucking crazy. As for exactly where the bottlenecks are and so on…

Dwarkesh Patel
Let's talk through the numbers here. You mentioned hundreds of millions of AI researchers. Right now, GPT-4o is like $15 for a million tokens outputted. A human thinks at 150 tokens a minute or something. If you do the math on that, for an hour's worth of human output, it's like $0.10 or something.

Leopold Aschenbrenner
It's cheaper than a human worker. It can't do the job yet.

Dwarkesh Patel
That's right. But by the time you're talking about models that are trained on the 10 GW cluster, then you have something that is four OOMs more expensive via inference, something like three OOMs. That's like $100/hour of labor. Now you're having hundreds of millions of such laborers. Is there enough compute to do this kind of labor with the model that is 1000 times bigger?

Leopold Aschenbrenner
Great question. I actually don't think inference costs for frontier models are necessarily going to go up that much.

Dwarkesh Patel
But isn't the test time sort of thing that it will go up even higher?

Leopold Aschenbrenner
We're just doing per token. Suppose each model token was the same as a human token thing at 100 tokens a minute. It'll use more, but the token calculation is already pricing that in. The question is per token pricing. GPT-3 when it launched was actually more expensive than GPT-4 now. Over vast increases in capability gains, inference cost has remained

constant. That's sort of wild, and it's worth appreciating. It gestures at an underlying pace of algorithmic progress.

There's a more theoretically grounded way to explain why inference costs would stay constant. On Chinchilla scaling laws, half of the additional compute you allocate to bigger models and half of it you allocate to more data. If we go with the basic story of 0.5 OOM/year more compute and 0.5 OOM/year of algorithmic progress you're saving 0.5 OOM/year. That would exactly compensate for making the model bigger.

The caveat is that obviously not all training efficiencies are also inference efficiencies. A bunch of the time they are. Separately, you can find inference efficiencies. Given this historical trend and baseline theoretical reason, it's not a crazy baseline assumption that the frontier models are not necessarily going to get more expensive, per token.

Dwarkesh Patel
Really? Okay, that's wild.

Leopold Aschenbrenner
We'll see. Even if they get 10x more expensive, then you have 10 million instead of 100 million. It's not really —

Dwarkesh Patel
But part of the intelligence explosion is that each of them has to run experiments that are GPT-4 sized. As a result, that takes up a lot of compute. Then you need to consolidate the results of experiments. What is the synthesized weight?

Leopold Aschenbrenner
You have much bigger inference compute anyway than your training. But the experiment compute is a constraint.

Dwarkesh Patel
Let’s go back to a more fundamental thing we're talking about here. In the series you say we should denominate the probability of getting to AGI in terms of OOMs of effective compute. Effective here accounts for the fact that there's a compute multiplier if you have a better algorithm. I'm not sure that it makes sense to be confident that this is a sensible way to project progress. It might be, but I have a lot of uncertainty about it.

It seems similar to somebody trying to project when we're going to get to the moon. They're looking at the Apollo program in the 1950s or something. They're like, "we have some amount of effective jet fuel and if we get more efficient engines, then we have more effective jet fuel. So we're going to determine the probability of getting to the moon based on the amount of effective jet fuel we have." I don't deny that jet fuel is important to launch rockets, but that seems like an odd way to denominate when you're going to get to the moon.

Leopold Aschenbrenner
I don't know how rocket science works, but I didn't get the impression that there's some clear scaling behavior with the amount of jet fuel. First of all, the scaling laws in AI have just held. A friend of mine pointed this out and it's a great point. If you look at the original Kaplan scaling laws paper — it went from 10^-9 to 10 petaflop days — and then concatenate additional compute from there to GPT-4, assuming some algorithmic progress, the scaling laws have held probably over 15 OOMs. It's a rough calculation so it's maybe even more. They've held for a lot of OOMs.

Dwarkesh Patel
They held for the specific loss function they're trained on, which is training the next token. Whereas the progress you are forecasting, we specifically know that that scaling can't work because of the data wall. There's some new thing that has to happen, and I'm not sure whether you can extrapolate that same scaling curve to tell us whether these hobblings will also be fixed. Is this not on the same graph?

Leopold Aschenbrenner
The hobblings are just a separate thing.

There's a few things here. On effective compute scaling, people center the scaling laws because they're easy to explain. Why does scaling matter?

The scaling laws came way after people, at least like Dario and Ilya, realized that scaling mattered. There's this great quote from Dario on your podcast. The models just want to learn. You make them bigger and they learn more. That's more important than the sort of loss curve.

That just applied across domains. You can look at this in benchmarks. Again, the headwind is the data wall. I'm bracketing that and talking about that separately.

The other thing is unhobblings. If you just put them on the effective compute graph, these unhobblings would be huge.

Dwarkesh Patel
What does it even mean? What is on the y-axis here?

Leopold Aschenbrenner
Say MLPR on this benchmark or whatever. We mentioned the LMSys differences, RLHF which is as good as 100x, chain-of-thought. Just going from this prompting change, a

simple algorithmic change can be like 10x effective compute increases on math benchmarks. This is useful to illustrate that unhobblings are large, but they're slightly separate things.

At a per token level, GPT-4 is not that far away from a token of my internal monologue. Even 3.5 to 4 took us from the bottom of the human range to the top of the human range on a lot of high school tests. It's a few more 3.5 to 4 jumps per token basis, per token intelligence. Then you've got to unlock the test time, solve the onboarding problem, make it use a computer, and then you're getting real close. The story might be wrong, but it is strikingly plausible.

The other thing I'll say is on the 2027 timeline, I do think it’s unlikely, but I do think there's worlds where there are AGI next year. That's basically if the test time compute overhang is really easy to crack. If it's really easy to crack, then you do like four OOMs of test time compute from a few hundred tokens to a few million tokens quickly. Then again, maybe it only takes one or two jumps equivalent equivalent to GPT-3.5 to 4, per token. One or two of those jumps per token plus test time compute and you basically have the proto automated engineer.

Dwarkesh Patel
I'm reminded of Steven Pinker’s book, The Better Angels of Our Nature. It talks about the secular decline in violence and war and everything. You can just plot the line from the end of World War Two. In fact from before World War Two, and then these are just aberrations. Basically as soon as it happens you get Ukraine, Gaza, etc.

Leopold Aschenbrenner
Impending ASI increasing crazy global conflict. ASI and crazy new WMDs.

Dwarkesh Patel
This is a thing that happens in history where you see a straight line then as soon as you make that prediction… Who is that famous author?

Leopold Aschenbrenner
Again, people have been predicting deep learning will hit a wall every year. Maybe one year they're right. But it's gone a long way and it hasn't hit a wall. You don't have that much more to go.

Dwarkesh Patel
This is a plausible story and let's just run with it and see what it implies.

In your series, you talk about alignment not from the perspective of “this is some doomer scheme to get the 0.01% of the probability distribution where things don't go off the rails.”

It's more about just controlling the systems and making sure they do what we intend them to do.

If that's the case, we're going to be in this sort of geopolitical conflict with China. What we're worried about is them making the CCP bots that go out and take the red flag of Mao across the galaxies. Shouldn't we then be worried about alignment as something that, in the wrong hands, enables brainwashing, and dictatorial control?

This seems like a worrying thing. This should be part of the sort of algorithmic secrets we keep hidden. The secret of how to align these models, because that's also something the CCP can use to control their models.

Leopold Aschenbrenner
In the world where you get the democratic coalition, yeah. Also, alignment is often dual use.

The alignment team developed RLHF and it was great. It was a big win for alignment, but it also obviously makes these models useful. So alignment enables the CCP bots. Alignment also is what you need to get the US AIs to follow the Constitution, disobey unlawful orders, and respect separation of powers and checks and balances. You need alignment for whatever you want to do. It's just the underlying technique.

Dwarkesh Patel
Tell me what you make of this take. I've been struggling with this a little bit.

Fundamentally, there's many different ways the future could go. There's one path that’s the Eliezer type: crazy AIs with nanobots take the future and turn everything into gray goo or paperclips.

The more you solve alignment, the more that path of the decision tree is circumscribed. The more you solve alignment, the more it's just different humans and the visions they have. Of course, we know from history that things don't turn out the way you expect. It's not like you can decide the future.

Leopold Aschenbrenner
That’s part of the beauty of it. You want these mechanisms like error correction —

Dwarkesh Patel
But from the perspective of anybody who's looking at the system it'll be like, “I can control where this thing is going to end up.” So the more you solve alignment — the more you circumscribe the different futures that are the result of AI will — the more that accentuates the conflict between humans and their visions of the future. The world where alignment is solved is the one in which you have the most sort of human conflict over where to take AI.

Leopold Aschenbrenner
By removing the worlds in which the AIs take over, the remaining worlds are the ones where the humans decide what happens. As we talked about, there are a whole lot of worlds there and how that could go.

Dwarkesh Patel
You think about alignment as just controlling these things. Just think a little forward. There are worlds in which hopefully human descendants, or some version of that in the future, merge with superintelligences. They have the rules of their own but they're in some sort of law and market-based order. I worry because you'll have things that are conscious and should be treated with rights. I'm thinking about what these alignment schemes actually are.

You read these books about what actually happened during the Cultural Revolution, what happened when Stalin took over Russia. You have very strong monitoring from different instances where everybody's tasked with watching each other. You have brainwashing. You have red teaming like the spy stuff you were talking about where you try to convince somebody you're a defector and you see if they defect with you. If they do, then you realize they're an enemy.

Maybe I'm stretching the analogy too far but the ease with which these alignment techniques actually map onto something you could have read about during Mao's Cultural Revolution is a little bit troubling.

Leopold Aschenbrenner
Sentient AI is a whole other topic. I don't know if we want to talk about it. I agree that it's going to be very important how we treat them. In terms of what you're actually programming these systems to do, again alignment is just a technical solution. It enables the CCP bots

Talking about checks and balances, the model is sort of like the Federal Reserve or Supreme Court justices. There's a funny way in which they're kind of this very dedicated order. It's amazing. They're actually quite high quality. They're really smart people who truly believe in and love the Constitution. They believe in their principles.

They have different persuasions, but they have very sincere debates about what is the meaning of the Constitution and what is the best actuation of these principles. By the way, I recommend SCOTUS oral arguments as the best podcast when I run out of high quality content on the Internet.

There's going to be a process of figuring out what the Constitution should be. This Constitution has worked for a long time. You start with that. Maybe eventually things change enough that you want edits to that. For example, on the checks and balances, they really love the Constitution. They believe in it and and they take it seriously.

At some point you are going to have AI police and AI military it'll be important to ensure that they believe in the Constitution the way that a Supreme Court justice does or the way that a Federal Reserve official takes their job really seriously.

The other important thing is that a bunch of different factions need their own AIs. It's really important that each political party gets to have their own superintelligence. You might totally disagree with their values, but it's important that they get to have their own kind of superintelligence. It's important that these classical liberal processes play out, including different people of different persuasions and so on. The AI advisors might not make them wise. They might not follow the advice or whatever, but it's important.

Dwarkesh Patel
You seem pretty optimistic about alignment. Let's get to the source of the optimism. You laid out different worlds in which we could get AI. There's one that you think has a low probability of happening next year, where GPT-5 plus scaffolding plus unhobblings gets you to AGI. There are also scenarios where it takes much longer.

GPT-4 seems pretty aligned in the sense that I don't expect it to go off the rails. Maybe with scaffolding, things might change. It looks pretty good, and maybe you will keep turning the cranks, and one of them gets you to ASI.

Is there any point at which the sharp left turn happens? Do you think it's plausible that when they start acting more like agents, this is something to worry about? Is there anything qualitative that you expect to change with regards to the alignment perspective?

Leopold Aschenbrenner
I don't know if I believe in this concept of a sharp left turn, but there are important qualitative changes that happen between now and somewhat superhuman systems early on in the intelligence explosion. There are also important qualitative changes that occur from early in the intelligence explosion to true superintelligence in all its power and might.

Let's talk about both of those. The first part of the problem is one we're going to have to solve ourselves. We have to align the initial AI and the intelligence explosion, the sort of automated Alec Radford. There are two important things that change from GPT-4. If you believe the story on synthetic data RL, self-play, to get past the data wall, and if you believe this unhobbling story, at the end you're going to have things that are agents. They'll do long-term planning. They have long horizons, which is a prerequisite to being able to do automated AI research.

Pre-training is alignment-neutral in the sense that it has good representations and representations of doing bad things, but it's not scheming against you. Misalignment can arise once you're doing more long-horizon training. For example, if you're training an AI to make money using reinforcement learning, it might learn to commit fraud, lie, deceive, or seek power simply because those are successful strategies in the real world. With RL, it explores, maybe it tries to hack and then it gets some money. If that's successful, that gets reward and that's just reinforced. There's more serious misalignments, like misaligned long-term goals, that necessarily have to be able to arise if you're able to get long-horizon systems.

Let's swap. What you want to do in that situation is add side constraints like "Don't lie.", "Don't deceive.", or "Don't commit fraud." How do you add those side constraints? The basic idea you might have is RLHF. You have this goal of making money, but you're watching what it's doing. If it starts trying to lie, deceive, commit fraud, or break the law, you give it a thumbs down and anti-reinforce that behavior.

The critical issue that arises is that these AI systems are becoming superhuman and will be able to do things that are too complex for humans to evaluate. Even early on in the intelligence explosion, the automated AI researchers and engineers might write millions, billions, or trillions of lines of complicated code. You won't understand what they're doing anymore. In those millions of lines of code, you don't know if it's hacking, exfiltrating itself, or trying to go for the nukes.

You don't know anymore. Thumbs up, thumbs down pure RLHF doesn't fully work anymore in this scenario. There's a hard technical problem of what do you do post-RLHF but it's a solvable problem. There's various things I'm bullish on. There's ways in which deep learning has shaped out favorably.

The second part of the picture is going from your initial systems in the intelligence explosion to superintelligence, many OOMs of improvement. By the end of it, you have a thing that's vastly smarter than humans. The intelligence explosion is really scary from an alignment point of view. If you have this rapid intelligence explosion in less than a year or two, you're going from systems where failure would be bad but not catastrophic to a world where if something goes awry, the AI could exfiltrate itself, start hacking the military, and do really bad things.

In less than a year, you're going from a world where the AI is some descendant of current systems that you understand and has good properties. It becomes something that potentially has a very alien and different architecture after having gone through another decade of ML advances. One salient example is legible and faithful chain-of-thought. A lot of the time when we're talking about these things, we're talking about how it has tokens of thinking and then uses many tokens of thinking. Maybe we bootstrap ourselves by pre-training it to learn to think in English, and then we do something else on top so it can do longer chains of thought.

It's very plausible to me that for the initial automated alignment researchers, we don't need to do any complicated mechanistic interpretability. You can just read what they're thinking, which is a huge advantage. However, it's very likely not the most efficient way to do it. There's probably some way to have a recurrent architecture with all internal states. That's a much more efficient way to do it.

That's what you get by the end of the year. You're going in this year from RLHF++ to something that's vastly superhuman. To us, it might be like an expert in the field compared to an elementary or middle school student. It's an incredibly hairy period for alignment. The thing you do have is the automated AI researchers. You can use them to also do alignment.

Dwarkesh Patel
In this world, why are we optimistic that the project is being run by the right people? Here's something to think about. OpenAI starts off with people who are very explicitly thinking about exactly these kinds of things.

Leopold Aschenbrenner
Yes but are they still there?

Dwarkesh Patel
No, but here's the thing. Even with the current leadership, you can find them in interviews and blog posts talking about it. You talked about what happens and it's not just you. Jan talked about it in his Tweet thread. There is some trade-off that has to be made with doing a flashy release this week and not next week because Google I/O is next week or whatever. The trade-off is made in favor of the more careless decision.

The government, the national security advisor or the military or whatever, is much less familiar with this kind of discourse. They're not like, "I'm worried the chain-of-thought is unfaithful. How do we think about the features that are represented here?" Why should we be optimistic that a project run by people like that will be thoughtful about these kinds of considerations?

Leopold Aschenbrenner
They might not be. Here's a few thoughts. First of all, the private world is extremely tough for alignment even if they nominally care. There's a couple of reasons. You have the race between the commercial labs. You don't have any headroom there to be like, aAh, actually we're going to hold back for three months, get this right. We're going to dedicate 90% of our compute to automated alignment research instead of just pushing the next OOM."

The other thing is that in the private world, China has stolen your weights. China has your secrets. They're right on your tails. You're in this fever struggle. There's no room at all for maneuver. It's absolutely essential to get alignment right. To get it during this intelligence explosion, to get it right, you need to have that room to maneuver and you need to have that clear lead. Again, maybe you've made the deal or whatever, but you're in an incredibly tough spot if you don't have this clear lead.

So the private world is kind of rough there. Whether people will take it seriously… I have some faith normal mechanisms of a liberal society. Wwe don't fully know yet if alignment is an issue. The science will develop. We're going to get better measurements of alignment. The case will be clear and obvious.

I worry that there's worlds where evidence is ambiguous. A lot of the most scary intelligence explosion scenarios are worlds in which evidence is ambiguous. But again, if evidence is ambiguous, then those are the worlds in which you really want the safety margins. Those are also the worlds in which running the intelligence explosion is sort of like running a war. The evidence is ambiguous. You have to make these really tough trade-offs. You better have a really good chain of command for that where they're not just YOLOing it.

Dwarkesh Patel
Let's talk a little about Germany. We're making the analogy to World War II. You made this. really interesting point many hours ago. Throughout history, World War Two is not unique at least when you think in proportion to the size of the population. Let's look at these other sorts of catastrophes where a substantial proportion of the population has been killed off.

After that, the nation recovers and they get back to their heights. What's interesting after World War Two is that Germany especially, and maybe Europe as a whole, they experienced vast economic growth in the immediate aftermath because of catch-up growth.

We're not talking about Germany potentially launching an intelligence explosion and them getting a seat at the AI table. We were talking about Iran and North Korea and Russia. We didn't talk about Germany.

Leopold Aschenbrenner
Because they're allies.

Dwarkesh Patel
So what happened? We had World War Two and now it didn't come back to the Seven Years' War or something.

Leopold Aschenbrenner
I'm generally very bearish on Germany. In this context, you're underrating a little bit. It's probably still one of the top five most important countries in the world. Europe overall still has a GDP that's close to the United States in size. There are things that Germany is actually kind of good at, like state capacity. The roads are good and they're clean and they're well-maintained.

In some sense, a lot of this is the flip side of things that are bad about Germany. In the US, there's a bit more of a Wild West feeling. It includes the kind of crazy bursts of creativity. It includes political candidates. There's a much broader spectrum. Both Obama and Trump are politicians you just wouldn't see in the much more confined kind of German political debate. I wrote a blog post at some point about this, "Europe's Political Stupor."

There's this punctilious sort of rule-following that is good in terms of keeping your state capacity functioning. But that is also a very constrained view of the world in some sense. After World War Two, there's a real backlash against anything elite. There are no elite high schools or elite colleges. Excellence isn't cherished.

Dwarkesh Patel
Why is that the logical, intellectual thing to rebel against if you're trying to overcorrect from the Nazis? Was it because the Nazis were very much into elitism? I don't understand why that's a logical sort of reaction.

Leopold Aschenbrenner
Maybe it was a counter reaction against the whole Aryan race and that sort of thing. Look at the end of World War I versus the end of World War II for Germany. A common narrative is that the Peace of Versailles was too strict on Germany. The peace imposed after World War Two was much more strict.

The whole country was destroyed. In most of the major cities, over half of the housing stock had been destroyed. In some birth cohorts, something like 40% of the men had died, Almost 20 million people displaced. It was huge and crazy.

Dwarkesh Patel
And the borders are way smaller than the Versailles borders.

Leopold Aschenbrenner
Yeah, exactly. There's also a complete imposition of a new political system on both sides. But in some sense, that worked out better than the post-World War I peace where there was this resurgence of German nationalism. In some sense, it's unclear if you want to wake the sleeping beast. At this point, it's gotten a bit too sleepy.

Dwarkesh Patel
It's an interesting point about how we underrate the American political system. I've been making the same correction myself. There was this book written by a Chinese economist called China's World View.

Overall, I wasn't a big fan, but it made a really interesting point, which was the way in which candidates rise up through the Chinese hierarchy for politics and administration. In some sense, it selects so that you're not going to get some Marjorie Taylor Greene or someone like that.

Leopold Aschenbrenner
Don't get that in Germany either.

Dwarkesh Patel
But he explicitly made the point in the book that it also means they're never going to get a Henry Kissinger or Barack Obama in China. By the time they end up in charge of the Politburo, they'll be some 60-year-old bureaucrat who's never ruffled any feathers.

Leopold Aschenbrenner
There's something really important about the very raucous political debate in the US. In general in America lots of people live in their own world. We live in this kind of bizarre little bubble in San Francisco and people. But that's important for the evolution of ideas, error correction, that sort of thing.

There are other ways in which the German system is more functional. There are also major mistakes, like with defense spending. Russia invades Ukraine and they're like, "wow, what did we do?"

Dwarkesh Patel
That's a really good point. The main issue is that everybody agrees.

Leopold Aschenbrenner
Exactly. There was no debate about it. It's a consensus Blob kind of thing.

On the China point, I have this experience of reading German newspapers and I would understand the German debate and the state of mind much more poorly without it from just afar. It is interesting just how impenetrable China is to me. It's a billion people.

Almost everything else is really globalized. You have a globalized Internet. I kind of have a sense what's happening in the UK. Even if I didn't read German newspapers, I would have a sense of what's happening in Germany. But I really don't feel like I have a sense of what is the state of mind, what is the state of political debate, of an average Chinese person or an average Chinese elite.

I find that distance kind of worrying. There are some people who do this and they do really great work where they go through the party documents and the party speeches. It seems to require a lot of interpretive ability. There are very specific words in Mandarin that mean one connotation, not the other connotation. It's interesting given how globalized everything is. Now we have basically perfect translation machines and it's still so impenetrable.

Dwarkesh Patel
That's really interesting. I'm sort of ashamed almost that I haven't done this yet. Many months ago when Alexey interviewed me on his YouTube channel, I said, "I'm meaning to go to China to actually see for myself what's going on." By the way, if anybody listening has a lot of context on China and if I went to China, could introduce me to people, please email me.

Leopold Aschenbrenner
You have to do some pods and find some of the Chinese AI researchers. It'd be so good. I don't know if they can speak freely.

Dwarkesh Patel
So they have these papers and on the paper they'll say who's a co-author. I was thinking of just cold emailing everybody, like, "Here's my Calendly. Let's just talk." I just want to see what the vibe is. Even if they don't tell me anything, I'm just like, "what kind of person is this? How westernized are they?"

I just remembered that, in fact, ByteDance, according to mutual friends we have at Google, cold emailed every single person on the Gemini paper and said, "if you come work for ByteDance, we'll make you a L8 engineer. You'll report directly to the CTO."

Leopold Aschenbrenner
That's how the secrets go over.

Dwarkesh Patel
I meant to ask this earlier. If there's only 100 or so people, maybe less, who are working on the key algorithmic secrets. If they hired one such person, is all the alpha that these labs have gone?

Leopold Aschenbrenner
If this person was intentional about it, they could get a lot. Actually, they could probably just exfiltrate the code. They could get a lot of the key ideas. Again up until recently, stuff was published but they could get a lot of the key ideas if they tried. There are a lot of people who

don't actually look around to see what the other teams are doing. But you kind of can. They could. It's scary.

Dwarkesh Patel
The project makes more sense there, where you can't just recruit a Manhattan Project engineer.

Leopold Aschenbrenner
These are secrets that can be used for probably every training run in the future. Maybe they’re the key to the data wall without which they can’t go on. They're going to give multipliers on compute worth hundreds of billions, trillions of dollars. All it takes is China to offer $100 million to somebody and say “come work for us.” I'm really uncertain on how seriously China is taking AGI right now.

One anecdote was related to me by another researcher in the field. They were at a conference with somebody, a Chinese AI researcher. He was talking to him and he was like, "I think it's really good that you're here. We have to have international coordination and stuff." Apparently this guy said that "I'm the most senior person that they're going to let leave the country to come to things like this."

Dwarkesh Patel
What's the takeaway?

Leopold Aschenbrenner
They're not letting really senior AI researchers leave the country. It’s a kind of classic Eastern Bloc move.

I don't know if this is true, but it’s what I heard.

Dwarkesh Patel
Let’s go back to the point you made earlier about being exposed to German newspapers. Earlier you mentioned you were interested in economics and law and national security. The variety in intellectual diet has exposed you to thinking about the geopolitical question here in ways that that others talking about AI aren’t.

This is the first episode I've done about this where we've talked about things like this. Now that I think about it, that’s weird given that this is an obvious thing in retrospect. I should have been thinking about it. That's one thing we've been missing.

What are you missing, not just in national security? What perspective are you probably underexposed to as a result? I guess you mentioned China as one.

Leopold Aschenbrenner
The China one is an important one. Another one would be a sort of very Tyler Cowen-esque take. You're not exposed to how a normal person in America will use AI. That kind of thing will be a bottleneck to the diffusion of these things. I'm overrating the revenue because I assume everyone is adopting it. But Joe Schmo engineer at a company, will they be able to integrate it? Also what’s the reaction to it? This was a question hours ago. Won’t people rebel against this? Will they not want to do the project? I don't know. Maybe they will.

Dwarkesh Patel
Here's a political reaction that I didn't anticipate. I already told you about this, but I'm just going to tell the story again. Tucker Carlson was recently on a Joe Rogan episode. They start talking about World War II.

Tucker says, "Well, listen. I'm going to say something that my fellow conservatives won't like, but I think nuclear weapons are immoral. I think it was obviously immoral that we use them on Nagasaki and Hiroshima."

Then he says, "In fact, nuclear weapons are always immoral, except when we would use them on data centers. In fact, it would be immoral not to use them on data centers, because, look, these people in Silicon Valley, these fucking nerds, are making superintelligence, and they say that it could enslave humanity. We made machines to serve humanity, not to enslave humanity. And they're just going on and making these machines. And so we should, of course, be nuking the data centers." That is definitely not a political reaction in 2024 I was expecting. It's going to be crazy.

Dwarkesh Patel
The thing we learned with COVID is also that the left-right reactions that you’d anticipate just based on hunches —

Leopold Aschenbrenner
It completely flipped multiple times. Initially the right was on it and the left was like, "This is racist." Then it flipped. The left was really into the lockdowns. The whole thing also is just so blunt and crude.

Probably in general, people like to make sort of complicated technocratic AI policy proposals. If things go kind of fairly rapidly on the path to AGI, there might not actually be that much space for complicated, clever proposals. It might just be much cruder reactions.

Dwarkesh Patel
You mentioned spies and national security getting involved and everything. You can talk about that in the abstract, but now that we're living in San Francisco we know many of the people who are doing the top AI research. It’s also a little scary to think about people I

personally know and am friends with. It's not unfeasible if they have secrets in their head that are worth $100 billion or something that you might see kidnapping, assassination, sabotage.

Leopold Aschenbrenner
Oh, their family. It's really bad. To the point on security, right now it's really foreign. At some point, as it becomes really serious, you're going to want the security guards.

Dwarkesh Patel
Presumably, you have thought about the fact that people in China will be listening to this and reading your series.

Somehow you made the trade-off that it's better to let the whole world know, including China, and wake them up to AGI than to stay silent. Part of the thing you're worried about is China waking up to AGI. I'm curious about that. Walk me through how you've thought about that trade-off.

Leopold Aschenbrenner
This is a tough trade-off. I thought about this a bunch. People in the PRC will read this.

To some extent the cat is out of the bag. AGI is a thing people are thinking about very seriously. That's not new anymore. A lot of these takes are kind of old or I had similar views a year ago. I might not have written it up a year ago, in part because I didn't think the cat wasn't out of the bag enough then.

To be able to manage this challenge, much broader swaths of society will need to wake up. If we're going to get the project, we actually need a broad bipartisan understanding of the challenges facing us. It's a tough trade-off. The need to wake up people in the United States, in the Western world, in the democratic coalition, is ultimately imperative. My hope is more people here will read it than in the PRC.

People sometimes underrate the importance of just kind of writing it up and laying out the strategic picture. You have done actually a great service to mankind in some sense with your podcast. It's overall been good.

Leopold Aschenbrenner
By the way, on the topic of Germany. We were talking at some point about immigration stories. You have an interesting story you haven't told, butI think you should tell it

Dwarkesh Patel
So a couple years ago, I was in college and I was 20. I was about to turn 21.

Leopold Aschenbrenner
You came from India when you were really young, right?

Dwarkesh Patel
Until I was eight or nine, I lived in India. Then we moved around all over the place. Because of the backlog for Indians we'd been in the queue for decades.

Leopold Aschenbrenner
Even though you came at eight, you're still on the H-1B.

Dwarkesh Patel
When you're 21 you get kicked off the queue and you have to restart the process. My dad's a doctor and I'm on his H-1B as a dependent. But when you're 21, you get kicked off. So I'm 20 and it just kind of dawns on me that this is my situation.

Leopold Aschenbrenner
You're completely screwed.

Dwarkesh Patel
I also had the experience with my dad. We moved all around the country. They have to prove, him being a doctor, that you can't get native talent.

Leopold Aschenbrenner
And you can't start a startup or anything. Even getting the H-1B for you would have been a 20% lottery, if you're lucky.

Dwarkesh Patel
Plus they had to prove that they can't get native talent, which meant that we lived in North Dakota for three years, West Virginia for three years, Maryland, West Texas.

So it dawned on me that this is my situation as I turn 21. I'll be on this lottery. Even if I get the lottery, I'll be a fucking code monkey for the rest of my life, because this thing isn't going to let up.

Leopold Aschenbrenner
Yeah. Can't do a startup.

Dwarkesh Patel
Exactly. At the same time, I had been reading for the last year and was super obsessed with Paul Graham essays. My plan at the time was to make a startup or something. I was super excited about that.

It just occurred to me that I couldn't do this. That just wasn't in the cards for me. I was kind of depressed about it. I remember I was in a daze through finals because it had just occurred to me. I was really anxious about it.

I remember thinking to myself at the time that if somehow I ended up getting my green card before I turned 21, there's no fucking way I'm becoming a code monkey. The feeling of dread that I have is this realization that I'm just going to have to be a code monkey. I realized that's my default path. If I hadn't made a proactive effort not to do that, I would have graduated college as a computer science student. I would have just done that. That's the thing I was super scared about. That was an important realization for me.

Anyway, COVID happened. Because of that, since there weren't any foreigners coming, the backlog got fast-tracked and by the skin of my teeth, like a few months before I turned 21, I ended up getting a green card for crazy, extremely contingent reasons.

Because I got a green card, I could —

Leopold Aschenbrenner
The whole podcast.

Dwarkesh Patel
I graduated college and I was bumming around. I graduated a semester early. I'm going to do this podcast and see what happens? If I didn't have a green card, I mean the best case scenario —

Leopold Aschenbrenner
It's such a cultural artifact. What is the impact of immigration reform? What is the impact of clearing 50,000 green cards in the backlog? You're such an amazing example how all of this is only possible contingent on that. It's just incredibly tragic that this is so dysfunctional.

Dwarkesh Patel
It's insane.

Leopold Aschenbrenner
I'm glad you did it. I'm glad you kind of tried the unusual path.

Dwarkesh Patel
I could only do it because I was extremely fortunate to get the green card. I had a little bit of saved up money. I got a small grant out of college, thanks to the Future Fund, to do this for like six months. It turned out really well. At each time, I was like, "Oh, okay. Podcast. Come on. I wasted a few months on this. Let's now go do something real." Something big would happen at each moment.

Leopold Aschenbrenner
You kept with it.

Dwarkesh Patel
There would always be something the moment I'm about to quit the podcast. Jeff Bezos would say something nice about me on Twitter. The Ilya episode gets like half a million views. Now this is my career. Looking back on it though, it was incredibly contingent that things worked out the right way.

Leopold Aschenbrenner
If the AGI stuff goes down, it'll be how most of the people who kind of end up feeling AGI first heard about it.

Dwarkesh Patel
You're also very linked with the story in many ways. I got like a $20,000 grant from Future Fund right out of college and that sustained me for six months or something. Without that…

Leopold Aschenbrenner
Tiny grant. It was kind of crazy. It goes to show how far small grants can go. Emergent Ventures, too.

Dwarkesh Patel
Exactly. Emergent Ventures. The last year I've been in San Francisco, we've just been in close contact the entire time and just bouncing ideas back and forth. People would be surprised by how much of the alpha I have I got from you, Sholto, Trenton and a couple others.

Leopold Aschenbrenner
It’s been an absolute pleasure.

Dwarkesh Patel
Likewise, it's been super fun. Here are some random questions for you. If you could convert to Mormonism and you could really believe it, would you do it? Would you push the button?

Leopold Aschenbrenner
Before I answer that question, one observation about the Mormons. There's an article that actually made a big impact on me. It was about the Mormons, by McKay Coppins in The Atlantic. He even interviewed Mitt Romney in it.

The thing he talked about was how the experience of growing up different, growing up very unusual, especially if you grew up Mormon outside of Utah. You’re the only person who doesn't drink caffeine, you don't drink alcohol, you're kind of weird. That got people prepared for being willing to be outside of the norm later on.

Mitt Romney was willing to take stands alone in his party because he believed what he believed was true. Probably not in the same way, but I feel a little bit like this from having grown up in Germany, having been kind of an outsider or something.

Growing up as an outsider gives you unusual strength later on to be willing to say what you think. So that is one thing I really appreciate about the Mormons, at least the ones that grow up outside of Utah.

The other thing is the fertility rates. They're good. They're important. They're going down as well. This is the thing that really clinched the fertility decline story for me. Even the Mormons.

Dwarkesh Patel
You're like, "Oh, this is like a good start. Mormons will replace everybody."

Leopold Aschenbrenner
I don't know if it's good, but at least some people will maintain high fertility rates. But no, even the Mormons. Once these religious subgroups that have high fertility rates grow big enough, they become too close in contact with normal society and become normalized. Their fertility rates drop from maybe like four to two in the course of 10-20 years.

People point to the Amish or whatever, but it's probably just not scalable. If you grow big enough, then there's just this overwhelming force of modernity that gets you.

No, if I could convert to Mormonism - look, I think there's something. I don't believe it, right? If I believed it, I obviously would convert to Mormonism, because you got to convert.

Dwarkesh Patel
But you can choose a world in which you do believe it.

Leopold Aschenbrenner
There's something really valuable in believing in something greater than yourself and having a certain amount of faith.

Dwarkesh Patel
You do, right? That's what your series is.

Leopold Aschenbrenner
It’s valuable to feel some sort of duty to something greater than yourself. Maybe my version of this is somewhat different. I feel some sort of duty to the historical weight on how this might play out. I feel some sort of duty to make that go well. I feel some sort of duty to our country, to the national security of the United States. We can be a force for a lot of good.

Dwarkesh Patel
Going back to OpenAI, there’s something that's especially impressive about that is. There are people at the company who have — through years and decades of building up savings from working in tech — probably tens of millions of dollars liquid and more than that in terms of their equity. Many people were concerned about the clusters and the Middle East and the secrets leaking to China and all these things.

The person who actually made a hassle about it — hassling people is so underrated — is the 22-year-old who has less than a year at the company, who doesn't have savings built up, who isn't a solidified member of the company.

Leopold Aschenbrenner
Maybe it's me being naive and not knowing how big companies work. Sometimes I'm a bit of a speech deontologist. I kind of believe in saying what you think. Sometimes friends tell me I should be more of a speech consequentialist.

Dwarkesh Patel
I mean I think about the amount of people who, when they have the opportunity to talk to the person, will just bring up the thing. I've been with you in multiple contexts. I guess I shouldn't reveal who the person is or what the context was.

I've just been very impressed that the dinner begins and by the end, somebody who has a major voice in how things go is seriously thinking about a worldview they would have found incredibly alien before the dinner. I've been impressed that you just give them the spiel and hassle them.

Leopold Aschenbrenner
I just feel this stuff pretty viscerally now. There was a time when I thought about this stuff a lot, but it was kind of like econ models and these theoretical abstractions. You talk about human brain size or whatever.

Since at least last year, I feel like I can see it. I feel it. I can sort of see the cluster that AGI can be trained on. I can see the kind of rough combination of algorithms and the people that will be involved and how this is going to play out. Look, we'll see how it plays out. There are many ways this could be wrong. There are many ways it could go, but this could get very real.

Dwarkesh Patel
Should we talk about what you're up to next?

Leopold Aschenbrenner
Sure, yeah.

Dwarkesh Patel
You're starting an investment firm with anchor investments from Nat Friedman, Daniel Gross, Patrick Collison, John Collison. First of all, why is this the thing to do if you believe AGI is coming in a few years? Why the investment firm?

Leopold Aschenbrenner
Good question. Fair question. A couple of things. We talked about this earlier, but the screen doesn't go blank when AGI intelligence happens. People really underrate the decade after you have the intelligence explosion. That's maybe the most wild period. The decade after is also going to be wild.

This combination of human institutions with superintelligence and crazy geopolitical things going on. You have this broadening of this explosive growth. Basically, it's going to be a really important period. Capital will really matter. Eventually we're going to go to the stars, going to go to the galaxies.

Part of the answer is just that done right, there's a lot of money to be made. If AGI were priced in tomorrow, you could maybe make 100x. Probably you can make even way more than that because of the sequencing and capital matters.

The other reason is just some amount of freedom and independence. There are some people who are very smart about this AI stuff and who see it coming. Almost all of them are constrained in various ways. They're in the labs, they're in some other position where they can't really talk about this stuff.

I've really admired the thing you've done. It's really important that there are voices of reason on this stuff publicly or people who are in positions to kind of advise important actors and so on.

Basically, this investment firm will be kind of like a brain trust on AI. It's going to be all about situational awareness. We're going to have the best situational awareness in the business. We're going to have way more situational awareness than any of the people who manage money in New York. We're definitely going to do great on investing, but it's the same sort of situational awareness that is going to be important for understanding what's happening, being a voice of reason publicly, and being able to be in a position to advise.

Dwarkesh Patel
The book about Peter Thiel, they had an interesting quote about his hedge fund. It got terrible returns. So this isn't the example...

Leopold Aschenbrenner
It blew up. That's sort of the bear case. It's too theoretical.

Dwarkesh Patel
They had an interesting quote that it's basically a think tank inside of a hedge fund.

Leopold Aschenbrenner
That's what I'm going to try to build.

Dwarkesh Patel
Presumably you've thought about the ways in which these kinds of things can blow up. There's a lot of interesting business history books about people who got the thesis right but timed it wrong. They buy into the idea that the Internet's going to be a big deal. They sell at the wrong time and buy at the wrong time during the dot-com boom. They miss out on the gains even though they're right about the. What is the trick to preventing that kind of thing?

Leopold Aschenbrenner
Obviously, not blowing up is task number one and two. This investment firm is going to just be betting on AGI. We're going to be betting on AGI and superintelligence before the decade is out, taking that seriously, making the bets you would make if you took that seriously. If that's wrong, the firm is not going to do that well.

The thing you have to be resistant to is you have to be able to resist one or a couple or a few individual calls. AI stagnates for a year because of the data wall, or you got the call wrong on when revenue would go up. That's pretty critical. You have to get the timing right. The sequence of bets on the way to AGI is actually pretty critical. People underrate it.

Where does the story start? Obviously, the only bet over the last year was Nvidia. It's obvious now, very few people did it. This is also a classic debate I and a friend had with another colleague of ours. This colleague was really into TSMC. He was just kind of like, "well, these fabs are going to be so valuable. With Nvidia, there's just a lot of idiosyncratic risk, right? Maybe somebody else will make better GPUs." That was basically right.

But only Nvidia had the AI beta, because only Nvidia was kind of like large fraction AI. The next few doublings would just meaningfully explode their revenue, whereas TSMC was a couple percent AI. Even though there's going to be a few doublings of AI, it was not going to make that big of an impact. The only place to find the AI beta, basically was Nvidia for a while.

Now it's broadening. Now TSMC is like 20% AI by 2027 or something. That's what they're saying. When we're doubling, it'll be kind of like a large fraction of what they're doing. There's a whole stack. There's people making memory and coops and power. Utilities companies are starting to get excited about AI. They're like, "Power production in the United States will grow not 2.5%, but 5% over the next five years." I'm like, "No, it'll grow more."

At some point, a Google or something becomes interesting. People are excited about them with AI because it's like, "oh, AI revenue will be $10 billion or tens of billions." I don't really care about them before then. I care about it once you get the AI beta. At some point Google will get $100 billion of revenue from AI. Probably their stock will explode. They're going to become a $5 trillion, $10 trillion company anyway.

The timing there is very important. You have to get the timing right. You have to get the sequence right. At some point, actually, there's going to be real headwind to equities from real interest rates. In these sorts of explosive growth worlds, you would expect real interest rates to go up a lot. On the supply side it'll be around the demand for money because people are going to be making these crazy investments, initially in clusters and then in the robo factories or whatever. They're going to be borrowing like crazy. They want all this capital, high ROI.

On the consumer saving side, to give up all this capital, it'll be the Euler equation, standard intertemporal transfer trade-off of consumption.

Dwarkesh Patel
Very standard.

Leopold Aschenbrenner
Some of our friends have a paper on this. Basically, if consumers expect real growth rates to be higher, interest rates are going to be higher because they're less willing to give up consumption today for consumption in the future.

At some point real interest rates will go up. Higher growth rate expectations mean equities go down because the interest rate effect outweighs the growth rate effect.

At some point there's the big bond short. You got to get that right. You got to get it right on nationalization. There's this whole sequence of things.

Dwarkesh Patel
And the unknown unknowns.

Leopold Aschenbrenner
Unknown unknowns, yeah. You've got to be really, really careful about your overall risk positioning. If you expect these crazy events to play out, there's going to be crazy things you didn't foresee.

You do also want to make the bets that are tailored to your scenarios in the sense of you want to find bets that are bets on the tails. I don't think anyone is expecting interest rates to go above 10%, real interest rates. There's at least a serious chance of that before the decade is out. Maybe there's some cheap insurance you can buy on that.

Dwarkesh Patel
Very silly question. In these worlds, are financial markets where you make these kinds of bets going to be respected? Is my Fidelity account going to mean anything when we have 50% economic growth? Who's like, "We have to respect his property rights"?

Leopold Aschenbrenner
That's pretty deep into it, the bond short, the 50% growth. That's pretty deep into it. Again, there's this whole sequence of things. I think property rights will be respected. At some point, there's going to be figuring out the property rights for the galaxies. That'll be interesting.

Dwarkesh Patel
That will be interesting. Going back to your strategy about how important the 2030s will be for how the rest of the future goes, you want to be in a position of influence by that point because of capital.

As far as I know, there's probably a whole bunch of literature on this, I'm just riffing. The landed gentry before the beginning of the Industrial Revolution, I'm not sure if they were able to leverage their position in a sort of Georgist or Piketty-type sense, in order to accrue the returns that were realized through the Industrial Revolution. I don't know what happened. At some point, they just weren't the landed gentry.

I'd be concerned that even if you make great investment calls, you'll be like the guy who owned a lot of farmland before the Industrial Revolution. The guy who's actually going to make a bunch of money is the one with the steam engine. Even he doesn't make that much money because most of the benefits are widely diffused and so forth.

Leopold Aschenbrenner
The analog is you sell your land and you put it all in the people who are building the new industries. The real depreciating asset for me is human capital. I was valedictorian of Columbia. The thing that made you special is you're smart. In four years, it might not matter because it's automatable.

A friend joked that the investment firm is perfectly hedged for me. Either AGI happens this decade and my human capital depreciates, but I turn it into financial capital, or no AGI happens and the firm doesn't do well, but I'm still in my twenties and smart.

Dwarkesh Patel
Excellent. What's your story for why AGI hasn't been priced in? Financial markets are supposed to be very efficient, so it's hard to get an edge. Naively, you might say, "I've looked at these scaling curves, and they imply we'll be buying much more compute and energy than analysts realize." Shouldn't those analysts be broke by now? What's going on?

Leopold Aschenbrenner
I used to believe in the EMH guy as an economist. But now, I think there are groups of smart people, like those in San Francisco, who have alpha over the rest of society in seeing the future.

It's like with COVID. A similar group of people saw it coming and called it completely corrected. They shorted the market and did really well. Why isn't AGI priced in? It's like asking why the government hasn't nationalized the labs yet. Society hasn't priced it in yet. It hasn't completely diffused. I might be wrong but not many people take these ideas seriously.

Dwarkesh Patel
There are a couple of other ideas I was playing around with that we haven't gotten to talk about yet. One's systems competition. One of my favorite books about World War II is Victor Davis Hanson's summary of everything. He explains why the Allies made better decisions than the Axis.

Leopold Aschenbrenner
Why did they?

Dwarkesh Patel
There were decisions the Axis made that were pretty good, like blitzkrieg.

Leopold Aschenbrenner
That was sort of by accident though.

Dwarkesh Patel
In what sense? That they just had the infrastructure left over?

Leopold Aschenbrenner
My read of it is that blitzkrieg wasn't an ingenious strategy. Their hand was forced. This is the very Adam Tooze-ian story of World War II. There's the concept of a long war versus a

short war, which is important. Germany realized that if they were in a long war, including the United States, they would not be able to compete industrially. Their only path to victory was to make it a short war. That worked much more spectacularly than they thought, allowing them to take over France and much of Europe.

The decision to invade the Soviet Union was related to the western front because they needed resources like oil. Auschwitz was actually a giant chemical plant to produce synthetic oil and other materials. It was the largest industrial project in Nazi Germany. They thought, “we crushed them in World War I, it’ll be easy. We’ll invade, get the resources, and then fight on the western front.” Even during the invasion of the Soviet Union, even though a large number of the deaths happened there, a large fraction of German industrial production—planes, naval forces, and so on—was directed towards the western front and the western allies.

By the way, this concept of a long war versus a short war is interesting, especially when thinking about the China competition. I worry about the decline of latent American industrial capacity. China builds like 200 times more ships than we do right now.

Maybe we have superiority in the non-AI world in military materiel and can win a short war or defend Taiwan. If it drags on, China might be better able to mobilize industrial resources in a way we can’t anymore. This is also relevant to AI. If building AGI requires a trillion-dollar cluster instead of a $100 billion cluster, or even if it’s on the $100 billion cluster, it really matters if you can do an order of magnitude more compute for your superintelligence. Maybe right now they’re behind, but they have the raw latent industrial capacity to outbuild us.

That matters both in the run-up to AGI and afterward. You have the superintelligence on your cluster, and then it’s time to expand the explosive growth. Will we let the robo-factories run wild? Maybe not, but maybe China will. How many drones will we produce? There’s an industrial explosion that I worry about.

Dwarkesh Patel
You’ve got to be one of the few people in the world who is both concerned about alignment but also wants to ensure we let the robo-factories proceed once we get ASI to beat out China. That’s very interesting.

Leopold Aschenbrenner
It’s all part of the picture.

Dwarkesh Patel
Speaking of ASIs and the robot factories and robo armies. One of the interesting things is the question of what you do with industrial-scale intelligence. Obviously, it's not chatbots. It's very hard to predict.

The history of oil is very interesting. In the 1860s, we figured out how to refine oil. A geologist discovered it, and then Standard Oil got started. There was a huge boom, changing American politics. Legislators were bought out by oil interests. Presidents were elected based on divisions about oil and breaking them up.

All this happened before the car was invented. The light bulb was invented 50 years after oil refining was discovered. Most of Standard Oil's history is before the car is invented. It was just kerosene lamps just used for lighting.

Leopold Aschenbrenner
So they thought oil would just no longer be relevant?

Dwarkesh Patel
Yeah. There was a concern that Standard Oil would go bankrupt when the light bulb was invented. You realize there's an immense amount of compressed energy here. You're going to have billions of gallons of this stuff a year. It's hard to predict in advance what you can do with that. Later on, it turns out it's used for transportation and cars.

With intelligence, maybe one answer is the intelligence explosion. But even after that, you have all these ASIs and enough compute, especially the compute they'll build to run —

Leopold Aschenbrenner
Hundreds of millions of GPUs will hum.

Dwarkesh Patel
What are we doing with that? It's very hard to predict in advance. It'll be very interesting to figure out what the Jupiter brains will be doing.

So there's situational awareness of where things stand now, and we've gotten a good dose of that. A lot of what we're talking about now couldn't have been predicted many years back. Part of your worldview implies that things will accelerate because of AI.

Many unpredictable factors will become evident over time, like how people, the political system, and foreign adversaries will react. Situational awareness isn't just knowing where things stand now, but being in a position to react appropriately to new information and to change your worldview and recommendations accordingly.

What is the appropriate way to think about situational awareness as a continuous process rather than as a one-time realization?

Leopold Aschenbrenner
This is great. There's a sort of mental flexibility and willingness to change your mind that's really important. This is how a lot of brains have been broken in the AGI debate. The doomers were prescient about AGI a decade ago, but they haven't updated on the empirical realities of deep learning. Their proposals are naive and unworkable. It doesn't really make sense.

Some people come in with a predefined ideology, like e/accs. They like to shitpost about technology but they're not actually thinking it through. You have stagnationists who think this stuff is just chatbots and not risky or those not considering the immense national security implications.

There's a risk of calcification of worldview when you publicly articulate a position and cling to it despite evidence against it. So I want to give a big disclaimer. It's valuable to paint a concrete and visceral picture. This is currently my best guess on how this decade will go. If it goes anything like this, it'll be wild. Given the rapid pace of progress, we're going to keep getting a lot more information and it's important to keep your head on straight.

I feel like the most important thing here is that. This relates to some of the stuff we talked about the world being surprisingly small. I used to think important things were being handled by capable people in government and AI labs.

From personal experience, and seeing how COVID was managed, I realized that not everyone is on it. There's not somebody else who's on it and making sure this goes well. What really matters is that good people take these issues as seriously as they deserve, have situational awareness, are willing to change their minds, and face reality head-on. I'm counting on those good people.

Dwarkesh Patel
Alright, that's a great place to close.