Dwarkesh Podcast #58 - Dario Amodei (Anthropic CEO) - The Hidden Pattern Behind Every AI Breakthrough

Published - August 8, 2023

Transcribed by - thepodtranscripts.com

Dwarkesh Patel
Today I have the pleasure of speaking with Dario Amodei, the CEO of Anthropic, and I'm really excited about this one.

Dario, thank you so much for coming on the podcast.

Dario Amodei
Thanks for having me.

Dwarkesh Patel
First question. You have been one of the very few people who has seen scaling coming for years. As somebody who's seen it coming, what is fundamentally the explanation for why scaling works? Why is the universe organized such that if you throw big blobs of compute at a wide enough distribution of data, the thing becomes intelligent?

Dario Amodei
I think the truth is that we still don't know. It's almost entirely an empirical fact. It's a fact that you could sense from the data and from a bunch of different places but we still don't have a satisfying explanation for it.

If I were to try to make one and I'm just kind of waving my hands when I say this, there's these ideas in physics around long tail or power law of correlations or effects. When a bunch of stuff happens, when you have a bunch of features, you get a lot of the data in the early fat part of the distribution before the tails. For language, this would be things like — "Oh, I figured out there are parts of speech and nouns follow verbs." And then there are these more and more subtle correlations.

So it kind of makes sense why every log or order of magnitude that you add, you capture more of the distribution. What's not clear at all is why does it scale so smoothly with parameters? Why does it scale so smoothly with the amount of data?

You can think up some explanations of why it's linear. The parameters are like a bucket, and the data is like water, and so size of the bucket is proportional to size of the water. But why does it lead to all this very smooth scaling? We still don't know. There's all these explanations. Our chief scientist, Jared Kaplan did some stuff on fractal manifold dimension that you can use to explain it.

So there's all kinds of ideas, but I feel like we just don't really know for sure.

Dwarkesh Patel
And by the way, for the audience who is trying to follow along. By scaling, we're referring to the fact that you can very predictably see how if you go from Claude-1 to Claude-2 that the loss in terms of whether it can predict the next token scales very smoothly.

Okay, so we don't know why it's happening, but can you at least predict empirically that here is the loss at which this ability will emerge, here is the place where this circuit will emerge? Is that at all predictable or are you just looking at the loss number?

Dario Amodei
That is much less predictable. What's predictable is this statistical average, this loss, this entropy. And it's super predictable. It's sometimes predictable even to several significant figures which you don't see outside of physics. You don't expect to see it in this messy empirical field. But specific abilities are actually very hard to predict. Back when I was working on GPT-2 and GPT-3, when does arithmetic come in place? When do models learn to code? Sometimes it's very abrupt.

It's like how you can predict statistical averages of the weather, but the weather on one particular day is very hard to predict.

Dwarkesh Patel
Dumb it down for me. I don't understand manifolds, but mechanistically, it doesn't know addition yet and suddenly now it knows addition. What has happened?

Dario Amodei
This is another question that we don't know the answer to. We're trying to answer this with things like mechanistic interpretability. You can think about these things like circuits snapping into place. Although there is some evidence that when you look at the models being able to add things, its chance of getting the right answer shoots up all of a sudden. But if you look at what's the probability of the right answer? You'll see it climb from like one in a million to one in 100,000 to one in a 1000 long before it actually gets the right answer. In many of these cases there's some continuous process going on behind the scenes. I don't understand it at all.

Dwarkesh Patel
Does that imply that the circuit or the process for doing addition was pre existing and it just got increased in salience?

Dario Amodei
I don't know if there's this circuit that's weak and getting stronger. I don't know if it's something that works, but not very well. I think we don't know and these are some of the questions we're trying to answer with mechanistic interpretability.

Dwarkesh Patel
Are there abilities that won't emerge with scale?

Dario Amodei
I definitely think that things like alignment and values are not guaranteed to emerge with scale. One way to think about it is you train the model and it's basically predicting the world, it's understanding the world. Its job is facts not values. It's trying to predict what comes next. But there's free variables here — What should you do? What should you think? What should you value? There aren't bits for that.

There's just — if I started with this I should finish with this. If I started with this other thing I should finish with this other thing. And so I think that's not going to emerge.

Dwarkesh Patel
If it turns out that scaling plateaus before we reach human level intelligence, looking back on it, what would be your explanation? What do you think is likely to be the case if that turns out to be the outcome?

Dario Amodei
I would distinguish some problem with the fundamental theory with some practical issue. One practical issue we could have is we could run out of data. For various reasons, I think that's not going to happen but if you look at it very naively we're not that far from running out of data. So it's like we just don't have the data to continue the scaling curves. Another way it could happen is we just use up all of the compute that was available and that wasn't enough and then progress is slow after that. I wouldn't bet on either of those things happening but they could.

From a fundamental perspective, I personally think it's very unlikely that the scaling laws will just stop. If they do, another reason could just be that we don't have quite the right architecture. If we tried to do it with an LSTM or an RNN the slope would be different. It still might be that we get there but there are some things that are just very hard to represent when you don't have the ability to attend far in the past that transformers have. If somehow we just hit a wall and it wasn't about the architecture I'd be very surprised by that. We're already at the point where to me the things the models can't do don't seem to be different in kind from the things they can do.

You could have made a case a few years ago that they can't reason, they can't program. You could have drawn boundaries and said maybe you'll hit a wall. I didn't think we would hit a wall, a few other people didn't think we would hit a wall, but it was a more plausible case then. It's a less plausible case now.

It could happen. This stuff is crazy. We could hit a wall tomorrow. If that happens my explanation would be there's something wrong with the loss when you train on next word prediction.

If you really want to learn to program at a really high level, it means you care about some tokens much more than others and they're rare enough that the loss function over focuses on the appearance, the things that are responsible for the most bits of entropy, and instead they don't focus on this stuff that's really essential. So you could have the signal drowned out in the noise. I don't think it's going to play out that way for a number of reasons. But if you told me — Yes, you trained your 2024 model. It was much bigger and it just wasn't any better, and you tried every architecture and didn't work, that's the explanation I would reach for.

Dwarkesh Patel
Is there a candidate for another loss function? If you had to abandon next token prediction.

Dario Amodei
I think then you would have to go for some kind of RL. There's many different kinds. There's RL from immune feedback, there's RL against an objective, there's things like Constitutional AI. There's things like amplification and debate. These are kind of both alignment methods and ways of training models.

You would have to try a bunch of things, but the focus would have to be on what do we actually care about the model doing? In a sense, we're a little bit lucky that predict the next word gets us all these other things we need. There's no guarantee.

Dwarkesh Patel
From your worldview it seems there's a multitude of different loss functions that it's just a matter of what can allow you to just throw a whole bunch of data at it. Next token prediction itself is not significant.

Dario Amodei
The thing with RL is you get slowed down a bit because you have to design how the loss function works by some method. The nice thing with the next token prediction is it's there for you. It's the easiest thing in the world. So I think it would slow you down if you couldn't scale in just that very simplest way.

Dwarkesh Patel
You mentioned that data is likely not to be the constraint. Why do you think that is the case?

Dario Amodei
There's various possibilities here and for a number of reasons I shouldn't go into the details, but there's many sources of data in the world and there's many ways that you can also generate data. My guess is that this will not be a blocker.

Maybe it would be better if it was, but it won't be.

Dwarkesh Patel
Are you talking about multimodal?

Dario Amodei
There's just many different ways to do it.

Dwarkesh Patel
How did you form your views on scaling? How far back can we go? And then you would be basically saying something similar to this.

Dario Amodei
This view that I have formed gradually from 2014 to 2017. My first experience with it was my first experience with AI. I saw some of the early stuff around AlexNet in 2012. I always had wanted to study intelligence but before I was just like, this doesn't seem like it's actually working. All the way back to 2005. I'd read Ray Kurzweil's work. I'd read even some of Eliezer's work on the early Internet back then. And I thought this stuff kind of looks far away. I look at the AI stuff of today and it's not anywhere close.

But with AlexNet I was like, oh, this stuff is actually starting to work. So I joined Andrew Ng's group at Baidu. I had been in a different field and this was my first experience with AI and it was a bit different from a lot of the academic style research that was going on elsewhere in the world.

I kind of got lucky in that the task that was given to me and the other folks there. It was just to make the best speech recognition system that you can.

There was a lot of data available, there were a lot of GPUs available. It posed the problem in a way that was amenable to discovering that kind of scaling was a solution. That's very different from being a postdoc whose job is to come up with an idea that seems clever and new and makes your mark as someone who's invented something.

I just tried the simplest experiments. I was just fiddling with some dials. I was like, try adding more layers to the RNN, try training it for longer, what happens? How long does it take to overfit? What if I add new data and repeat it less times? And I just saw these very consistent patterns.

I didn't really know that this was unusual or that others weren't thinking in this way. This was almost like beginner's luck. It was my first experience with it and I didn't really think about it beyond speech recognition. I was just like, oh, I don't know anything about this field. There are zillions of things people do with machine learning. But I'm like, weird, this seems to be true in the speech recognition field.

It was just before OpenAI started that I met Ilya, who you interviewed. One of the first things he said to me was — "Look. The models, they just want to learn. You have to understand this. The models, they just want to learn." And it was a bit like a Zen Koan. I listened to this and I became enlightened.

And over the years, I would be the one who would formalize a lot of these things and kind of put them together, but what that told me was that the phenomenon that I'd seen wasn't just some random thing. It was broad. It was more general. The models just want to learn. You get the obstacles out of their way. You give them good data, you give them enough space to operate in, you don't do something stupid like condition them badly numerically, and they want to learn. They'll do it.

Dwarkesh Patel
What I find really interesting about what you said is there were many people who were aware that these things are really good at speech recognition or at playing these constrained games. Very few extrapolated from there like you and Ilya did to something that is generally intelligent.

What was different about the way you were thinking about it versus how others were thinking about it? What made you think it's getting better at speech in this consistent way, it will get better at everything in this consistent way.

Dario Amodei
I genuinely don't know. At first when I saw it for speech, I assumed this was just true for speech or for this narrow class of models. I think it was just that over the period between 2014 and 2017, I tried it for a lot of things and saw the same thing over and over again. I watched the same being true with Dota. I watched the same being true with robotics. Many people thought that as a counterexample, but I just thought, well, it's hard to get data for robotics, but if we look within the data that we have, we see the same patterns.

I think people were very focused on solving the problem in front of them. It's very hard to explain why one person thinks one way and another person thinks a different way. People just see it through a different lens. They are looking vertically instead of horizontally. They're not thinking about the scaling, they're thinking about how do I solve my problem? And for robotics, there's not enough data. That can easily abstract to — scaling doesn't work because we don't have the data.

For some reason, and it may just have been random, I was obsessed with that particular direction.

Dwarkesh Patel
When did it become obvious to you that language is the means to just feed a bunch of data into these things? Or was it just you ran out of other things. Like robotics, there's not enough data. This other thing, there's not enough data.

Dario Amodei
I think this whole idea of the next word prediction, that you could do self supervised learning, together with the idea that there's so much richness and structure there for predicting the next word. It might say two plus two equals and you have to know the answer is four. It might be telling the story about a character.

Basically, it's posing to the model the equivalent of these developmental tests that get posed to children. Mary walks into the room and puts an item in there and then Chuck walks into the room and removes the item and Mary doesn't see it. What does Mary think?

To get this right in the service of predicting the next word the models are going to have to solve all these theory of mind problems, solve all these math problems. And so my thinking was just, well, you scale it up as much as you can. There's kind of no limit to it.

And I think I kind of abstractly had that view but the thing that really solidified and convinced me was the work that Alec Radford did on GPT-1. Which was that not only could you get this language model that could predict things very well but you could also fine tune it. In those days, you needed to fine tune it to do all these other tasks.

So I was like, wow, this isn't just some narrow thing where you get the language model right. It's sort of halfway to everywhere. You get the language model right and then with a little move in this direction, it can solve this logical dereference test or whatever. And with this other thing, it can solve translation or something. And then you're like, wow, I think there's really something to do. And of course, we can really scale it.

Dwarkesh Patel
One thing that's confusing, or that would have been hard to see — If you told me in 2018 we'll have models in 2023, like Claude 2 that can write theorems in the style of Shakespeare, whatever theory you want, they can ace standardized test with open ended questions, just all kinds of really impressive things, I would have said — Oh, you have AGI. You clearly have something that is human level intelligence.

While these things are impressive, it clearly seems we're not at human level, at least in the current generation and potentially for generations to come. What explains this discrepancy

between super impressive performance in these benchmarks and the things you could describe versus general intelligence?

Dario Amodei
That was one area where actually I was not prescient and I was surprised as well.

When I first looked at GPT-3 and the kind of things that we built in the early days at Anthropic, my general sense was that it seems like they've really grasped the essence of language. I'm not sure how much we need to scale them up. Maybe what's more needed from here is like RL and all the other stuff.

In 2020 I thought we can scale this a bunch more but I wonder if it's more efficient to scale it more or to start adding on these other objectives like RL. I thought maybe if you do as much RL as you've done pre training for a 2020 style model, that's the way to go.

Scaling it up will keep working. But is that really the best path? And I don't know, it just keeps going. I thought it had understood a lot of the essence of language but then there's further to go.

Stepping back from it. One of the reasons why I'm sort of very empiricist about AI, about safety, about organizations, is that you often get surprised. I feel like I've been right about some things but still with these theoretical pictures ahead, been wrong about most things. Being right about 10% of the stuff sets you head and shoulders above many people.
If you look back to these diagrams that are like, here's the village idiot, here's Einstein. Here's the scale of intelligence. And the village idiot and Einstein are very close to each other.

Maybe that's still true in some abstract sense or something but it's not really what we're seeing, is it? We're seeing that it seems like the human range is pretty broad and we don't hit the human range in the same place or at the same time for different tasks.

Like, write a sonnet in the style of Cormac McCarthy. I'm not very creative, so I couldn't do that but that's a pretty high level human skill. And even the model is starting to get good at stuff like constrained writing like, write a page about X without using the letter E.

I think the models might be superhuman or close to superhuman at that. But when it comes to proving relatively simple mathematical theorems, they're just starting to do the beginning of it. They make really dumb mistakes sometimes and they really lack any kind of broad correcting your errors or doing some extended task.

So it turns out that intelligence isn't a spectrum. There are a bunch of different areas of domain expertise. There are a bunch of different kinds of skills. Memory is different. It's all

formed in the blob, it's not complicated. But to the extent it even is on the spectrum, the spectrum is also wide.

If you asked me ten years ago, that's not what I would have expected at all, but I think that's very much the way it's turned out.

Dwarkesh Patel
Oh, man. I have so many questions just as a follow up on that.

Do you expect that given the distribution of training that these models get from massive amounts of internet data versus what humans got from evolution, that the repertoire of skills that elicits will be just barely overlapping? Will it be like concentric circles? How do you think about that? Do those matter?

Dario Amodei
Clearly there's certainly a large amount of overlap because a lot of the things these models do have business applications and many of their business applications are doing things that are helping humans to be more effective at things. So the overlap is quite large.

If you think of all the activity that humans put on the internet in text, that covers a lot of it, but it probably doesn't cover some things. Like the models learn a physical model of the world to some extent, but they certainly don't learn how to actually move around in the world. Again, maybe that's easy to fine tune.

So there are some things that the models don't learn that humans do. And then the models also learn things that humans don't, for example, to speak fluent Base 64. I don't know about you, but I never learned that.

Dwarkesh Patel
How likely do you think it is that these models will be superhuman for many years at economically valuable tasks while they are still below humans in many other relevant tasks that prevents an intelligence explosion or something?

Dario Amodei
This kind of stuff is really hard to know so I'll give that caveat. You can kind of predict the basic scaling laws and then this more granular stuff, which we really want to know to know how this all is going to go, is much harder to know.

My guess would be the scaling laws are going to continue. Again, subject to — do people slow down for safety or for regulatory reasons? But let's just put all that aside and say we have the economic capability to keep scaling. If we did that, what would happen?

My view is we're going to keep getting better across the board and I don't see any area where the models are super, super weak or not starting to make progress. That used to be true of math and programming, but over the last six months the 2023 generation of models, compared to the 2022 generation, has started to learn that. There may be more subtle things we don't know. And so I kind of suspect, even if it isn't quite even, that the rising tide will lift all the boats.

Dwarkesh Patel
Does that include the thing you were mentioning earlier where if there's an extended task, it loses its train of thought or its ability to just execute a series of steps?

Dario Amodei
That's going to depend on things like RL training to have the model do longer horizon tasks. I don't expect that to require a substantial amount of additional compute. I think that was probably an artifact of thinking about RL in the wrong way and underestimating how much the model had learned on its own.

In terms of are we going to be superhuman in some areas and not others? I think it's complicated. I could imagine that we won't be superhuman in some areas because they involve embodiment in the physical world. And then what happens? Do the AIs help us train faster AIs? And those faster AIs wrap around and solve that? Do you not need the physical world? It depends what you mean. Are we worried about an alignment disaster? Are we worried about misuse, like making weapons of mass destruction? Are we worried about AI taking over research from humans? Are we worried about it reaching some threshold of economic productivity where it can do what the average human does? I think these different thresholds have different answers, although I suspect they will all come within a few years.

Dwarkesh Patel
Let me ask about those thresholds. If Claude was an employee at Anthropic, what salary would it be worth? Is it meaningfully speeding up AI progress?

Dario Amodei
It feels to me like an intern in most areas, but then some specific areas where it's better than that.

One thing that makes the comparison hard is that the form factor is not the same as a human. If you were to behave like one of these chat bots, I guess we could have this conversation, but they're more designed to answer single or a few questions. They don't have the concept of having a long life of prior experience. We're talking here about things that I've experienced in the past and chat bots don't have that.

There's all kinds of stuff missing and so it's hard to make a comparison. They feel like interns in some areas and then they have areas where they spike and are really savants, where they may be better than anyone here.

Dwarkesh Patel
But does the overall picture of something like an intelligence explosion make sense to you? My former guest, Carl Shulman, has this very detailed model of an intelligence explosion. As somebody who would actually see that happening, does that make sense to you? As they go from interns to entry level software engineers. Those entry level software engineers increase your productivity...

Dario Amodei
I think the idea that as AI systems become more productive, first they speed up the productivity of humans, then they equal the productivity of humans, and then in some meaningful sense are the main contributor to scientific progress that happens at some point. That basic logic seems likely to me although I have a suspicion that when we actually go into the details, it's going to be weird and different than we expect. That in all the detailed models, we're thinking about the wrong things or we're right about one thing, and then are wrong about ten other things. I think we might end up in a weirder world than we expect.

Dwarkesh Patel
When you add all this together, what does your estimate of when we get something kind of human level look like?

Dario Amodei
It depends on the thresholds. In terms of someone looks at the model and even if you talk to it for an hour or so, it's basically like a generally well educated human, that could be not very far away at all. I think that could happen in two or three years.

The main thing that would stop it would be if we hit certain safety thresholds and stuff like that. So if a company or the industry decides to slow down or we're able to get the government to institute restrictions that moderate the rate of progress for safety reasons, that would be the main reason it wouldn't happen. But if you just look at the logistical and economic ability to scale, we're not very far at all from that.

Now that may not be the threshold where the models are existentially dangerous. In fact, I suspect it's not quite there yet. It may not be the threshold where the models can take over most AI research. It may not be the threshold where the models seriously change how the economy works.

I think it gets a little murky after that and all of those thresholds may happen at various times after that. But in terms of the base technical capability of — it kind of sounds like a reasonably generally educated human across the board. I think that could be quite close.

Dwarkesh Patel
Why would it be the case that it could pass a Turing Test for an educated person but not be able to contribute or substitute for human involvement in the economy?

Dario Amodei
A couple of reasons. One is just that the threshold of skill isn't high enough, comparative advantage. It doesn't matter that I have someone who's better than the average human at every task. What I really need for AI research is to find something that is strong enough to substantially accelerate the labor of the thousand experts who are best at it. We might reach a point where the comparative advantage of these systems is not great.

Another thing that could be the case is that there are these mysterious frictions that don't show up in naive economic models but you see it whenever you go to a customer or something. You're like — "Hey, I have this cool chat bot." In principle, it can do everything that your customer service bot does or this part of your company does, but the actual friction of how do we slot it in? How do we make it work? That includes both just the question of how it works in a human sense within the company, how things happen in the economy and overcome frictions, and also just, what is the workflow? How do you actually interact with it?

It's very different to say, here's a chat bot that looks like it's doing this task or helping the human to do some task as it is to say, okay, this thing is deployed and 100,000 people are using it.

Right now lots of folks are rushing to deploy these systems but in many cases, they're not using them anywhere close to the most efficient way that they could. Not because they're not smart, but because it takes time to work these things out. And so I think when things are changing this fast, there are going to be all of these frictions.

These are messy realities that don't quite get captured in the model. I don't think it changes the basic picture. I don't think it changes the idea that we're building up this snowball of, the models help the models get better and can accelerate what the humans do. And eventually it's mostly the models doing the work.

You zoom out far enough that's happening. But I'm skeptical of any kind of precise mathematical or exponential prediction of how it's going to be. I think it's all going to be a mess. But what we know is it's on a metaphorical exponential, and it's going to happen fast.

Dwarkesh Patel
How do those different exponentials which we've been talking about net out?

One was the scaling laws themselves are power laws with decaying marginal loss parameter or something. The other exponential you talked about is, these things can get involved in the process of AI research itself, speeding it up.

Those two are sort of opposing exponentials. Does it net out to be superlinear or sublinear? And also you mentioned that the distribution of intelligence might just be broader. After we get to this point in two to three years, what does that look like?

Dario Amodei
I think it's very unclear. We're already at the point where if you look at the loss, the scaling laws are starting to bend. We've seen that in published model cards offered by multiple companies. So that's not a secret at all.

But as they start to bend, each little bit of entropy of accurate prediction becomes more important. Maybe these last little bits of entropy are the difference between a physics paper as Einstein would have written it as opposed to some other physicist.

It's hard to assess significance from this. It certainly looks like in terms of practical performance, the metrics keep going up relatively linearly, although they're always unpredictable. It's hard to see that.

And then the thing that I think is driving the most acceleration is just more and more money is going into the field. People are seeing that there's just a huge amount of economic value and so I expect the price, the amount of money spent on the largest models, to go up by like a factor of 100 or something. And for that to then be concatenated with the chips are getting faster, the algorithms are getting better because there's so many people working on this now.

Again, I'm not making a normative statement here. This is what should happen. I'm not even saying this necessarily will happen because there's important safety and government questions here which we're very actively working on. I'm just saying, left to itself, this is what the economy is going to do.

Dwarkesh Patel
We'll get to those questions in a second. But how do you think about the contribution of Anthropic to that increase in the scope of this industry. There's an argument you can make that, with that investment, we can work on safety stuff at Anthropic, another that says you're raising the salience of this field in general.

Dario Amodei
It's all costs and benefits. The costs are not zero. A mature way to think about these things is not to deny that there are any costs, but to think about what the costs are and what the benefits are. I think we've been relatively responsible in the sense that we didn't cause the big acceleration that happened late last year and at the beginning of this year. We weren't the ones who did that.

And honestly, if you look at the reaction of Google, that might be ten times more important than anything else. And then once it had happened, once the ecosystem had changed, then we did a lot of things to stay on the frontier.

It's like any other question. You're trying to do the things that have the lowest costs and the biggest benefits and that causes you to have different strategies at different times.

Dwarkesh Patel
One question I had for you while we were talking about the intelligence stuff was, as a scientist yourself, what do you make of the fact that these things have basically the entire corpus of human knowledge memorized and they haven't been able to make a single new connection that has led to a discovery?

Whereas if even a moderately intelligent person had this much stuff memorized, they would notice — Oh, this thing causes this symptom. This other thing also causes this symptom. There's a medical cure right here.

Shouldn't we be expecting that kind of stuff?

Dario Amodei
I'm not sure. These words. Discovery. Creativity. One of the lessons I've learned is that in the big blob of compute, these ideas often end up being fuzzy and elusive and hard to track down.

But I think there is something here. The models do display a kind of ordinary creativity. Things like, write a sonnet in the style of Cormac McCarthy or Barbie. There is some creativity to that and they do draw new connections of the kind that an ordinary person would draw.

I agree with you that there haven't been any “big” scientific discoveries. I think that's a mix of just the model skill level is not high enough yet. I was on a podcast last week where the host said, “I don't know, I play with these models. They're kind of mid. They get a B or a B minus.”

That is going to change with the scaling.

I do think there's an interesting point about, well, the models have an advantage, which is they know a lot more than us. Shouldn't they have an advantage already, even if their skill level isn't quite high? Maybe that's kind of what you're getting at.

I don't really have an answer to that. It seems certainly like memorization and facts and drawing connections is an area where the models are ahead. And I do think maybe you need those connections and you need a fairly high level of skill.

Particularly in the area of biology, for better and for worse, the complexity of biology is such that the current models know a lot of things right now and that's what you need to make discoveries and draw connections. It's not like physics where you need to think and come up with a formula. In biology you need to know a lot of things. and so I do think the models know a lot of things and they have a skill level that's not quite high enough to put them together.

I think they are just on the cusp of being able to put these things together.

Dwarkesh Patel
On that point. Last week in your Senate testimony, you said that these models are two to three years away from potentially enabling large scale bio terrorism attacks. Can you make that more concrete without obviously giving the kind of information that would result in speeding that up? Is it one shotting how to weaponize something or do you have to fine tune an open source model? What would that actually look like?

Dario Amodei
I think it'd be good to clarify this because we did a blog post and the Senate testimony and various people didn't understand the point or didn't understand what we'd done.

Today you can ask the models all kinds of things about biology and get them to say all kinds of scary things, but often those scary things are things that you could Google, and I'm therefore not particularly worried about that. I think it's actually an impediment to seeing the real danger, where someone just says — Oh, I asked this model to tell me some things about smallpox, and it will.

That is actually not what I'm worried about. We spent about six months working with folks who are the most expert in the world on how do biological attacks happen, what would you need to conduct such an attack, and how do we defend against such an attack?

They worked very intensively on just the entire workflow of trying to do a bad thing. It's not one shot, it's a long process. There are many steps to it. It's not just like I asked the model for this one page of information. And again, without going into any detail, the thing I said in the Senate testimony is, there are some steps where you can just get information on Google. There are some steps that are what I'd call missing. They're scattered across a

bunch of textbooks, or they're not in any textbook. They're kind of implicit knowledge, and they're not explicit knowledge. They're more like, I have to do this lab protocol, and what if I get it wrong? Oh, if this happens, then my temperature was too low. If that happened, I needed to add more of this particular reagent.

What we found is that for the most part, those key missing pieces, the models can't do them yet, but we found that sometimes they can, and when they can, sometimes they still hallucinate, which is the thing that's keeping us safe. But we saw enough signs of the models doing those key things well. And if we look at state of the art models and go backwards to previous models, we look at the trend, it shows every sign that two or three years from now, we're going to have a real problem.

Dwarkesh Patel
Yeah, especially the thing you mentioned on the log scale. You go from one in 100 times, it gets it right, to one in ten, to..

Dario Amodei
Exactly. I've seen many of these “groks” in my life. I was there when I watched when GPT-3 learned to do arithmetic, when GPT-2 learned to do regression a little bit above chance, when with Claude we got better on all these tests of helpful, honest, harmless. I've seen a lot of groks. This is unfortunately not one that I'm excited about, but I believe it's happening.

Dwarkesh Patel
Somebody might say, listen, you were a co-author on this post that OpenAI released about GPT-2 where they said, we're not going to release the weights or the details here because we're worried that this model will be used for something bad. And looking back on it now, it's laughable to think that GPT-2 could have done anything bad. Are we just way too worried? This is a concern that doesn't make sense?

Dario Amodei
It is interesting. It might be worth looking back at the actual text of that post. I don't remember it exactly but it's still up on the Internet. It says something like, we're choosing not to release the weights because of concerns about misuse. But it also said, this is an experiment. We're not sure if this is necessary or the right thing to do at this time, but we'd like to establish a norm of thinking carefully about these things. You could think of it a little like the Asilomar conference in the 1970s where they were just figuring out recombinant DNA. It was not necessarily the case that someone could do something really bad with recombinant DNA. It's just the possibilities were starting to become clear. Those words, at least, were the right attitude.

Now I think there's a separate thing that people don't just judge the post, they judge the organization. Is this an organization that produces a lot of hype or that has credibility or something like that? And so that had some effect on it. I guess you could also ask, is it inevitable that you can't get across any message more complicated than this thing right here is dangerous.

You can argue about those but I think the basic thing that was in my head and the head of others who were involved in that, and what is evident in the post is, we actually don't know. We have pretty wide error bars on what's dangerous and what's not so we want to establish a norm of being careful.

By the way we have enormously more evidence now. We've seen enormously more of these groks now and so we're well calibrated but there's still uncertainty. In all these statements I've said, in two or three years we might be there. There's a substantial risk of it and we don't want to take that risk. But I wouldn't say it's 100%. It could be 50-50.

Dwarkesh Patel
Okay, let's talk about cybersecurity, which in addition to bio risk is another thing Anthropic has been emphasizing. How have you avoided the cloud microarchitecture from leaking? Because, as you know, your competitors have been less successful at this kind of security.

Dario Amodei
Can't comment on anyone else's security, don't know what's going on in there. A thing that we have done is, there are these architectural innovations that make training more efficient. We call them compute multipliers because they're the equivalent of having more compute.

I don't want to say too much about our compute multipliers because it could allow an adversary to counteract our measures but we limit the number of people who are aware of a given compute multiplier to those who need to know about it.

So there's a very small number of people who could leak all of these secrets. There's a larger number of people who could leak one of them. But this is the standard compartmentalization strategy that's used in the intelligence community or resistance cells or whatever. Over the last few months we've implemented these measures. I don't want to jinx anything by saying, oh, this could never happen to us but I think it would be harder for it to happen. I don't want to go into any more detail.

By the way I'd encourage all the other companies to do this as well. As much as competitors architecture's leaking is narrowly helpful to Anthropic, it's not good for anyone in the long run. Security around this stuff is really important.

Dwarkesh Patel
Could you, with your current security, prevent a dedicated state level actor from getting the Claude 2 weights?

Dario Amodei
It depends how dedicated. Our head of security, who used to work on security for Chrome, which is a very widely used and attacked application, he likes to think about it in terms of — how much would it cost to attack Anthropic successfully? Again, I don't want to go into super detail of how much I think it will cost to attack and it's just inviting people. One of our goals is that it costs more to attack Anthropic than it costs to just train your own model. It doesn't guarantee things because, of course you need the talent as well so you might still, but attacks have risks, the diplomatic costs, and they use up the very sparse resources that nation state actors might have in order to do the attacks.

We're not there yet by the way. But I think we are at a very high standard of security compared to the size of company that we are. If you look at security for most 150 person companies there's just no comparison. But could we resist if it was a state actor's top priority to steal our model weights? No. They would succeed.

Dwarkesh Patel
How long does that stay true? Because at some point the value keeps increasing and increasing. And another part of this question is what kind of a secret is how to train Claude 3 or Claude 2?

For example, with nuclear weapons we had lots of spies. You just take a blueprint of the implosion device across and that's what you need. Is it more tacit here like the thing you were talking about with biology? You need to know how these reagents work or is it just like you got the blueprint, you got the microarchitecture and the hyperparameters?

Dario Amodei
There are some things that are like a one line equation and there are other things that are more complicated. I think compartmentalization is the best way to do it. Just limit the number of people who know about something. If you're a 1000 person company and everyone knows every secret, one, I guarantee you have a leaker and two, I guarantee you have a spy.

Dwarkesh Patel
Okay, let's talk about alignment and let's talk about mechanistic interpretability, which is the branch you guys specialize in. While you're answering this question, you might want to explain what mechanistic interpretability is.

The broader question is mechanistically, what is alignment? Is it that you're locking in the model into a benevolent character? Are you disabling deceptive circuits and procedures? What concretely is happening when you align a model?

Dario Amodei
As with most things, when we actually train a model to be aligned, we don't know what happens inside the model. There are different ways of training it to be aligned but we don't really know what happens. All the current methods that involve some kind of fine tuning of course have the property that the underlying knowledge and abilities that we might be worried about don't disappear. The model is just taught not to output them. I don't know if that's a fatal flaw or if that's just the way things have to be. I don't know what's going on inside mechanistically and I think that's the whole point of mechanistic interpretability. To really understand what's going on inside the models at the level of individual circuits.

Dwarkesh Patel
Eventually when it's solved, what does the solution look like? What is the case where if you're Claude 4, you do the mechanistic interpretability thing and you're like, I'm satisfied, it's aligned. What is it that you've seen?

Dario Amodei
We don't know enough to know that yet. I can give you a sketch for what the process looks like as opposed to what the final result looks like. Verifiability is a lot of the challenge here. We have all these methods that purport to align AI systems and do succeed at doing so for today's tasks.

But then the question is always if you had a more powerful model or if you had a model in a different situation, would it be aligned? This problem would be much easier if you had an oracle that could just scan a model and say okay, I know this model is aligned, I know what it'll do in every situation.

I think the closest thing we have to that is something like mechanistic interpretability. It's not anywhere near up to the task yet. But I guess I would say I think of it as almost like an extended training set and an extended test set. Everything we're doing, all the alignment methods we're doing are the training set. You can run tests in them, but will it really work out a distribution? Will it really work in another situation?

Mechanistic interpretability is the only thing that even in principle is the thing where it's more like an X-ray of the model than modification of the model. It's more like an assessment than an intervention. Somehow we need to get into a dynamic where we have an extended test set, an extended training set, which is all these alignment methods, and an extended test set which is kind of like you X-ray the model and say, okay, what worked and what didn't? In a way that goes beyond just the empirical test that you've run, where you're saying,

what is the model going to do in these situations? What is within its capabilities to do instead of, what did it do phenomenologically?

And of course we have to be careful about that. One of the things I think is very important is we should never train for interpretability because that's taking away that advantage. You even have the problem similar to validation versus test set, where if you look at the X-ray too many times, you can interfere. We should worry about that, but that's a much weaker process, it's not automated optimization. We should just make sure, as with validation and test sets, that we don't look at the validation set too many times before running the test set. But again, that's manual pressure rather than automated pressure.

So some solution where we have some dynamic between the training and test set where we're trying things out and we really figure out if they work via a way of testing them, that the model isn't optimizing against, some orthogonal way.

I think we're never going to have a guarantee, but some process where we do those things together. Some way to put extended training for alignment ability with extended testing for alignment ability together in a way that actually works. And not in a stupid way, there's lots of stupid ways to do this where you fool yourself.

Dwarkesh Patel
I still don't feel like I understand the intuition for why you think this is likely to work or this is promising to pursue. Let me ask the question in a more specific way, and excuse the tortured analogy.

If you're an economist and you want to understand the economy, you send a whole bunch of microeconomists out there. One of them studies how the restaurant business works. One of them studies how the tourism business works, one of them studies how the baking business works. And at the end, they all come together and you still don't know whether there's going to be a recession in five years or not.

Why is this not like that? Where you have an understanding of how induction heads work in a two layer transformer, we understand modular arithmetic. How does this add up to — Does this model want to kill us? What does this model fundamentally want?

Dario Amodei
A few things on that. That's the right set of questions to ask. I think what we're hoping for in the end is not that we'll understand every detail, but again, I would give the X-ray or the MRI analogy. We can be in a position where we can look at the broad features of the model and say, is this a model whose internal state and plans are very different from what it externally represents itself to do? Is this a model where we're uncomfortable that far too much of its

computational power is devoted to doing what look like fairly destructive and manipulative things?

We don't know for sure whether that's possible, but at least some positive signs that it might be possible. Again, the model is not intentionally hiding from you, it might turn out that the training process hides it from you. I can think of cases where if the model is really super intelligent, it thinks in a way so that it affects its own cognition. We should think about that, we should consider everything. I suspect that it may roughly work to think of the model as if it's trained in the normal way, just getting to above human level. It may be a reasonable assumption, you should check, that the internal structure of the model is not intentionally optimizing against us.

I'd give an analogy to humans. It's actually possible to look at an MRI of someone and predict above random chance whether they're a psychopath. There was actually a story a few years back about a neuroscientist who was studying this, and then he looked at his own scan and discovered that he was a psychopath and then everyone in his life was like — No, this is obvious. You're a complete asshole. You must be a psychopath. And he was totally unaware of this.

The basic idea that there can be these macro features, psychopath is probably a good analogy for it, this is what we would be afraid of, a model that's charming on the surface, very goal oriented, and very dark on the inside. On the surface, their behavior might look like the behavior of someone else, but their goals are very different.

Dwarkesh Patel
A question somebody might have is, you're trying to empirically estimate if these activations are suspicious but is this something we can afford to be empirical about? Or do we need a very good first principal theoretical reason to think — No, it's not just that these MRIs of the model correlate with being bad. We need just some deep rooted math proof that this is aligned.

Dario Amodei
It depends what you mean by empirical. A better term would be phenomenological. I don't think we should be purely phenomenological in like, here are some brain scans of really dangerous models and here are some other brain scans. The whole idea of mechanistic interpretability is to look at the underlying principles and circuits.

But I guess the way I'd think about it is like, on one hand, I've actually always been a fan of studying these circuits at the lowest level of detail that we possibly can. And the reason for that is that's kind of how you build up knowledge. Even if you're ultimately aiming for there's too many of these features, it's too complicated. At the end of the day, we're trying to build something broad and we're trying to build some broad understanding. I think the way you

build that up is by trying to make a lot of these very specific discoveries. You have to understand the building blocks and then you have to figure out how to use that to draw these broad conclusions even if you're not going to figure out everything.

You should probably talk to Chris Olah, who would have much more detail. He controls the interpretability agenda. He's the one who decides what to do on interpretability. This is my high level thinking about it, which is not going to be as good as his.

Dwarkesh Patel
Does the bull case on Anthropic rely on the fact that mechanistic interpretability is helpful for capabilities?

Dario Amodei
I don't think so at all. I think in principle it's possible that mechanistic interpretability could be helpful with capabilities. We might, for various reasons, not choose to talk about it if that were the case.

That wasn't something that I or any of us thought of at the time of Anthropic's founding. We thought of ourselves as people who are good at scaling models and good at doing safety on top of those models. We think that we have a very high talent density of folks who are good at that. My view has always been talent density beats talent mass. That's more of our bullcase. Talent density beats talent mass.

I don't think it depends on some particular thing. Others are starting to do mechanistic interpretability now, and I'm very glad that they are. A part of our theory of change is paradoxically to make other organizations more like us.

Dwarkesh Patel
I'm sure talent density is important but another thing Anthropic has emphasized is that you need to have frontier models in order to do safety research. And of course, actually be a company as well.

Somebody might guess that the current frontier models, GPT-4, Claude 2 cost one hundred million dollars or something like that...

Dario Amodei
That general order of magnitude in very broad terms is not wrong.

Dwarkesh Patel
But two to three years from now, the kinds of things you're talking about, we're talking more and more orders of magnitude to keep up with that. If it's the case that safety requires us to

be on the frontier, what is a case in which Anthropic is competing with these leviathans to stay on that same scale?

Dario Amodei
It's a situation with a lot of trade offs. It's not easy. Maybe I'll just answer the questions one by one.

To go back to why is safety so tied to scale? Some people don't think it is. But if I just look at what have been the areas where safety methods have been put into practice or worked for something, for anything, even if we don't think they'll work in general.

I go back to thinking of all the ideas, something like debate and amplification. Back in 2018 when we wrote papers about those at OpenAI, it was like, human feedback isn't quite going to work, but debate and amplification will take us beyond that. But then if you actually look at the attempts to do debates, we're really limited by the quality of the model. For two models to have a debate that is coherent enough that a human can judge it so that the training process can actually work, you need models that are at or maybe even beyond on some topics the current frontier. You can come up with the method, you can come up with the idea without being on the frontier but for me, that's a very small fraction of what needs to be done. It's very easy to come up with these methods. It's very easy to come up with, oh, the problem is X, maybe a solution is Y.

I really want to know whether things work in practice, even for the systems we have today, and I want to know what kinds of things go wrong with them. I just feel like you discover ten new ideas and ten new ways that things are going to go wrong by trying these in practice. I think that empirical learning is just not as widely understood as it should be.

I would say the same thing about methods like constitutional AI, and some people say, oh, it doesn't matter. We know this method doesn't work, it won't work for pure alignment. I neither agree nor disagree with that. I think that's just kind of overconfident. The way we discover new things and understand the structure of what's going to work and what's not is by playing around with things. Not that we should just blindly say, oh, this worked here, and so it'll work there. But you really start to understand the patterns, like with the scaling laws.

Even mechanistic interpretability, which might be the one area I see where a lot of progress has been made without the frontier models, we're seeing in the work that OpenAI put out a couple months ago, that using very powerful models to help you auto interpret the weak models. Again, that's not everything you can do in interpretability, but that's a big component of it and we found it useful too.

So you see this phenomenon over and over again where the scaling and the safety are these two snakes that are coiled with each other, always even more than you think. Even with

interpretability, three years ago, I didn't think that this would be as true of interpretability, but somehow it manages to be true. Why? Because intelligence is useful. It's useful for a number of tasks. One of the tasks it's useful for is figuring out how to judge and evaluate other intelligence and maybe someday even for doing the alignment research itself.

Dwarkesh Patel
Given all that's true, what does that imply for Anthropic when in two to three years, these leviathans are doing like $10 billion training runs?

Dario Amodei
Choice one is if we can't, or if it costs too much to stay on the frontier, then we shouldn't do it and we won't work with the most advanced models, we'll see what we can get with models that are not quite as advanced. You can get some non zero value there but I'm skeptical that the value is all that high or the learning can be fast enough to really be in favor of the task.

The second option is you just find a way. You just accept the trade offs. And the trade offs are more positive than they appear because of a phenomenon that I've called Race to the Top. I could go into that later, but let me put that aside for now.

And the third phenomenon is that as things get to that scale, it may coincide with starting to get into some non trivial probability of very serious danger. I think it's going to come first from misuse, the biorisk stuff that I talked about. I don't think we have the level of autonomy yet to worry about some of the alignment stuff happening in two years, but it might not be very far behind that at all. That may lead to unilateral or multilateral or government enforced decisions not to scale as fast as we could, which we support. That may end up being the right thing to do. I hope things go in that direction, and then we don't have this hard trade off between we're not in the frontier and can't quite do the research as well as we want or influence other orgs as well as we want, or versus we're on the frontier and have to accept the trade-offs which are net positive, but have a lot in both directions.

Dwarkesh Patel
On the misuse versus misalignment, those are both problems as you mentioned but in the long scheme of things, say 30 years down the line, which do you think will be considered a bigger problem?

Dario Amodei
I think it's going to be much less than 30 years. I'm worried about both. If you have a model that could in theory, take over the world on its own, if you were able to control that model, then it follows pretty simply that if a model was following the wishes of some small subset of people and not others, then those people could use it to take over the world on their behalf. The very premise of misalignment means that we should be worried about misuse as well, with similar levels of consequences.

Dwarkesh Patel
But some people who might be more doomery than you would say — you're already working towards the optimistic scenario there because you've at least figured out how to align the model with the bad guys. Now you just need to make sure that it's aligned with the good guys instead.

Why do you think that you could get to the point where it's aligned with the bad guys? You haven't already solved this.

Dario Amodei
I guess if you had the view that alignment is completely unsolvable, then you'd be like, well, we're dead anyway so I don't want to worry about misuse. That's not my position at all.

But also you should think in terms of what's a plan that would actually succeed that would make things good. Any plan that actually succeeds, regardless of how hard misalignment is to solve, is going to need to solve misuse as well as misalignment.

As the AI models get better faster and faster, they're going to create a big problem around the balance of power between countries. They're going to create a big problem around, is it possible for a single individual to do something bad that it's hard for everyone else to stop? Any actual solution that leads to a good future needs to solve those problems as well. If your perspective is, we're screwed because we can't solve the first problem, so don't worry about problems two and three, that's not really a statement. You should worry about problems two and three. They're in our path no matter what.

Dwarkesh Patel
Yeah. In the scenario we succeed we have to solve all of them.

Dario Amodei
We should be planning for success not for failure.

Dwarkesh Patel
If misuse doesn't happen and the right people have the superhuman models, what does that look like? Who are the right people? Who is actually controlling the model five years from now?

Dario Amodei
My view is that these things are powerful enough that I think it's going to involve substantial involvement of some kind of government or assembly of government bodies. There are very naive versions of this. I don't think we should just hand the model over to the UN or whoever happens to be in office at a given time. I could see that going poorly. But it's too powerful. There needs to be some kind of legitimate process for managing this technology, which

includes the role of the people building it, includes the role of democratically elected authorities, includes the role of all the individuals who will be affected by it. At the end of the day, there needs to be some politically legitimate process.

Dwarkesh Patel
But what does that look like? If it's not the case that you just hand it to whoever the President is at the time, what does the body look like?

Dario Amodei
It's really hard to know these things ahead of time. People love to propose these broad plans and say, oh, this is the way we should do it. The honest fact is that we're figuring this out as we go along. I think we should try things and experiment with them with less powerful versions of the technology. We need to figure this out in time. But also it's not really the kind of thing you can know in advance.

Dwarkesh Patel
The long term benefit trust that you have. How would that interface with this body? Is that the body itself?

Dario Amodei
I think that the long term benefit trust is a much narrower thing. This is something that makes decisions for Anthropic. This is basically a body. It was described in a recent Vox article. We'll be saying more about it later this year. But it's basically a body that over time gains the ability to appoint the majority of the board seats of Anthropic. It's a mixture of experts in AI alignment, national security, and philanthropy in general.

Dwarkesh Patel
If Anthropic has AGI and if control of Anthropic is handed to them, doesn't that imply that control of AGI itself is handed to them?

Dario Amodei
That doesn't imply that Anthropic or any other entity should be the entity that makes decisions about AGI on behalf of humanity. I would think of those as different things. If Anthropic does play a broad role, then you'd want to widen that body to a whole bunch of different people from around the world. Or maybe you construe this as very narrow, and then there's some broad committee somewhere that manages all the AGIs of all the companies on behalf of anyone.

I don't know. I think my view is you shouldn't be overly constructive and utopian. We're dealing with a new problem here. We need to start thinking now about what are the governmental bodies and structures that could deal with it.

Dwarkesh Patel
Okay, so let's forget about governance. Let's just talk about what this going well looks like.

Obviously, there are things we can all agree on: cure all the diseases, solve all the fraud – things all humans would say, 'I'm down for that.' But now it's 2030. You've solved all the real problems that everybody can agree on. What happens next? What are we doing with a superhuman God?

Dario Amodei
I actually want to disagree with the framing of something like this. I get nervous when someone says, what are you going to do with a superhuman AI? We've learned a lot of things over the last 150 years about markets and democracy, and each person can define for themselves what the best way for them to have the human experience is, and that societies work out norms and what they value just in this very complex and decentralized way.

If you have these safety problems that can be a reason why there needs to be a certain amount of centralized control from the government until we've solved these problems.

But as a matter of — we've solved all the problems, now how do we make things good? I think most people, most groups, most ideologies that started with, let's sit down and think over what the definition of the good life is, have led to disaster.

Dwarkesh Patel
But this vision you have of a sort of tolerant, liberal, democracy, market oriented system with AGI. Each person has their own AGI? What does that mean?

Dario Amodei
I don't know. I don't know what it looks like. I guess what I'm saying is we need to solve the important safety problems and the important externalities. Those could be just narrowly about alignment, there could be a bunch of economic issues that are super complicated and that we can't solve. Subject to that, we should think about what's worked in the past. And in general, unitary visions for what it means to live a good life have not worked out well at all.

Dwarkesh Patel
On the opposite end of things going well or good actors having control of AI. We might want to touch on China as a potential actor in the space.

First of all, being at Baidu and seeing progress in AI happening generally, why do you think the Chinese have underperformed? Baidu had a scaling laws group many years back. Or is the premise wrong and I'm just not aware of the progress that's happening there?

Dario Amodei
The scaling laws group, that was an offshoot of the stuff we did with speech so there were still some people there but that was a mostly Americanized lab. I was there for a year. That was my first foray into deep learning. It was led by Andrew Ng. I never went to China. It was like a US lab. That was somewhat disconnected, although it was an attempt by a Chinese entity to kind of get into the game.

Since then I think they've maybe been very commercially focused and not as focused on these fundamental research side of things around scaling laws. I do think because of all the excitement with the release of ChatGPT in November or so, that's been a starting gun for them as well. And they're trying very aggressively to catch up now.

I think the US is substantially ahead but they're trying very hard to catch up now.

Dwarkesh Patel
How do you think China thinks about AGI? Are they thinking about safety and misuse or not?

Dario Amodei
I don't really have a sense. One concern I would have are people saying things like, China isn't going to develop an AI because they like stability or they're going to have all these restrictions to make sure things are in line with what the CCP wants. That might be true in the short term and for consumer products. My worry is that if the basic incentives are about national security and power, that's going to become clear sooner or later. If they see this as a source of national power, they're going to at least try to do what's most effective, and that could lead them in the direction of AGI.

Dwarkesh Patel
Assume they just get your blueprints or your code base or something, is it possible for them to spin up their own lab that is competitive at the frontier with the leading American companies?

Dario Amodei
I don't know about fast but I'm concerned about this. This is one reason why we're focusing so hard on cybersecurity. We've worked with our cloud providers. We had this blog post out about security where we said we have a two key system for access to the model weights. We have other measures that we put in place or are thinking of putting in place that we haven't announced. We don't want an adversary to know about them, but we're happy to talk about them broadly.

By the way all this stuff we're doing is not sufficient yet for a super determined state level actor at all. I think it will defend against most attacks and against a state level actor who's

less determined. But there's a lot more we need to do, and some of it may require new research on how to do security.

Dwarkesh Patel
Let's talk about what it would take at that point. We're at Anthropic offices and it's got good security. We had to get badges and everything to come in here. But what does the eventual version of this building or bunker or whatever where the AGI is built look like? Is it a building in the middle of San Francisco or are you out in the middle of Nevada or Arizona? What is a point in which you're Los Alamos-ing it?

Dario Amodei
At one point there was a running joke somewhere that the way building AGI would look like is, there would be a data center next to a nuclear power plant next to a bunker, and that we'd all kind of live in the bunker and everything would be local so it wouldn't get on the Internet.

If we take the rate at which all this is going to happen seriously, which I can't be sure of, then it does make me think that something like that might happen, but maybe not something quite as cartoonish.

Dwarkesh Patel
What is the timescale on which you think alignment is solvable? If these models are getting to human level in some things in two to three years, what is the point at which they're aligned?

Dario Amodei
This is a really difficult question because I actually think often people are thinking about alignment in the wrong way. There's a general feeling that it's like models are misaligned or there's like an alignment problem to solve. Like, someday we'll crack the Riemann hypothesis. I don't quite think it's like that. Not in a way that's worse or better. It might be just as bad or just as unpredictable.

When I think of why am I scared, there's a few things I think of — One is, the thing that's really hard to argue with is: There will be powerful models. They will be agentic. We're getting towards them. If such a model wanted to wreak havoc and destroy humanity or whatever, we have basically no ability to stop it. If that's not true, at some point we will reach the point where it's true as we scale the models. So that definitely seems to be the case.

A second thing that seems to be the case is that we seem to be bad at controlling the models. Not in any particular way, but they're just statistical systems and you can ask them a million things and they can say a million things and reply. And you might not have thought of a millionth and one thing that does something crazy. Or when you train them, you train them in this very abstract way and you might not understand all the consequences of what

they do in response to that. The best example we've seen of that is Bing and Sydney. I don't know how they trained that model. I don't know what they did to make it do all this weird stuff like threaten people and have this weird obsessive personality. But what it shows is that we can get something very different from and maybe opposite to what we intended.

I actually think fact number one and fact number two are enough to be really worried. You don't need all this detailed stuff about convergent instrumental goals or analogies to evolution. One and two for me are pretty motivated. Okay, this thing's going to be powerful. It could destroy us. And all the ones we've built so far are at pretty decent risk of doing some random shit we don't understand.

Dwarkesh Patel
If you say that we're going to get something with bioweapons or something that could be dangerous in two to three years, does the research agenda you have of mechanistic interpretability, constitutional AI and other RLHF stuff meaningfully contribute to preventing that in two to three years?

Dario Amodei
People talk about doom by default or alignment by default. I think it might be kind of statistical. With the current models, you might get Bing or Sydney or you might get Claude. If we take our current understanding and move that to very powerful models, you might just be in this world where you make something and depending on the details, maybe it's totally fine. Not really alignment by default, but just depends on a lot of the details. If you're very careful about all those details and you know what you're doing, you're getting it right but we have a high susceptibility to, you mess something up in a way that you didn't really understand was connected to something else. Actually, instead of making all the humans happy, it wants to turn them into pumpkins, just some weird shit. Because the models are so powerful, they're like these giants that are standing in a landscape and if they start to move their arms around randomly, they could just break everything.

I'm starting it with that kind of framing because I don't think we're aligned by default, I don't think we're doomed by default and have some problem we need to solve. It has some kind of different character.

Now what I do think is that hopefully within a timescale of two to three years we get better at diagnosing when the models are good and when they're bad. We get better at increasing our repertoire of methods to train the model that they're less likely to do bad things and more likely to do good things in a way that isn't just relevant to the current models but scales. And we can help develop that with interpretability as the test set. I don't think of it as, oh, man, we tried RLHF, it didn't work. We tried Constitutional AI, it didn't work. We tried this other thing, it didn't work. We tried mechanistic interpretability. Now we're going to try

something else. I think this frame of like, man, we haven't cracked the problem yet, we haven't solved the Riemann hypothesis isn't quite right.

Already with today's systems, we are not very good at controlling them and the consequences of that could be very bad. We just need to get more ways of increasing the likelihood that we can control our models and understand what's going on in them. And we have some of them so far. They aren't that good yet. But I don't think of this as binary. It works or it does not work. We're going to develop more. And I do think that over the next two to three years we're going to start eating that probability mass of ways things can go wrong. It's like in the core safety views paper, there's a probability mass of how hard the problem is.

I feel like that way of stating it isn't really even quite right because I don't feel like it's the Riemann hypothesis to solve. It's almost like right now if I try and juggle five balls or something, I can juggle three balls, I actually can, but I can't juggle five balls at all. You have to practice a lot to do that. If I were to do that, I would almost certainly drop them. And then just over time, you just get better at the task of controlling the balls.

Dwarkesh Patel
On that post in particular, what is your personal probability distribution? For the audience, the three possibilities are: One, it is trivial to align these models with RLHF++. Two, it is a difficult problem, but one that a big company could solve. Three, something that is basically impossible for human civilization currently to solve. If I'm capturing those three, What is your probability distribution over those three?

Dario Amodei
I'm not super into questions like what's your probability distribution of X? I think all of those have enough likelihood that they should be considered seriously. The question I'm much more interested in is, what could we learn that shifts probability mass between them?

Dwarkesh Patel
What is the answer to that?

Dario Amodei
I think that one of the things mechanistic interpretability is going to do more than necessarily solve problems is, it's going to tell us what's going on when we try to align models. It's basically going to teach us about this. One way I could imagine concluding that things are very difficult is if mechanistic interpretability sort of shows us that problems tend to get moved around instead of being stamped out or that, you get rid of one problem, you create another one. Or it might inspire us or give us insight into why problems are persistent or hard to eradicate or crop up.

For me to really believe some of these stories about, oh, there's always this convergent goal in this particular direction. I think the abstract story is not uncompelling, but I don't find it really compelling either, nor do I find it necessary to motivate all the safety work.

But the kind of thing that would really be like, oh man, we can't solve this is like, we see it happening inside the X-ray. I think right now there's way too many assumptions, there's way too much overconfidence about how all this is going to go. I have a substantial probability mass on — this all goes wrong, it's a complete disaster, but in a completely different way than anyone had anticipated it would.

Dwarkesh Patel
It would be beside the point to ask how it could go different than anyone anticipated.

On this, in particular, what information would be relevant? How much would the difficulty of aligning Claude 3 and the next generation of models be? Is that a big piece of information?

Dario Amodei
I think the people who are most worried are predicting that all the subhuman AI models are going to be alignable, They're going to seem aligned. They're going to deceive us in some way. It certainly gives us some information but I am more interested in what mechanistic interpretability can tell us because, again, you see this X ray, it would be too strong to say it doesn't lie, but at least in the current systems, it doesn't feel like it's optimizing against us. There are exotic ways that it could. I don't think anything is a safe bet here, but it's the closest we're going to get to something that isn't actively optimizing against us.

Dwarkesh Patel
Let's talk about the specific methods other than mechanistic interpretability that you guys are researching. When we talk about RLHF or Constitution AI, if you had to put it in terms of human psychology, what is the change that is happening? Are we creating new drives, new goals, new thoughts? How is the model changing in terms of psychology?

Dario Amodei
All those terms are inadequate for describing what's happening. It's not clear how useful they are as abstractions for humans either. I think we don't have the language to describe what's going on. And again, I'd love to have the X-ray. I'd love to look inside and say and kind of actually know what we're talking about instead of basically making up words, which is what I do what you're doing in asking this question. We should just be honest. We really have very little idea what we're talking about. It would be great to say, well, what we actually mean by that is this circuit within here turns on, and after we've trained the model, then this circuit is no longer operative or weaker. It's going to take a lot of work to be able to do that.

Dwarkesh Patel
Model organisms, which you hinted at before when you said we're doing these evaluations to see if they're capable of doing dangerous things now and currently not, how worried are you about a lab leak scenario? Where in fine tuning it or in trying to get these models to elicit dangerous behaviors, make bioweapons or something, you leak somehow and it actually makes the bioweapons instead of telling you it can make the bioweapons.

Dario Amodei
It's not that much of a concern with today's passive models. If we were to fine tune a model, we would do it privately and we work with the experts and so the leak would be like, suppose the model got open sourced or something. For now, it's mostly a security issue.

In terms of models truly being dangerous, we do have to worry that if we make a truly powerful model and we're trying to see what makes it dangerous or safe, then there could be more of a one shot thing where there's some risk that the model takes over. The main way to control that is to make sure that the capabilities of the model that we test are not such that they're capable of doing this.

Dwarkesh Patel
At what point would the capabilities be so high where you say, I don't even want to test this?

Dario Amodei
Well, there's different things. There's capability testing..

Dwarkesh Patel
But that itself could lead to... If you're testing replicate, what if it actually does?

Dario Amodei
Sure. But I think what you want to do is you want to extrapolate. We've talked with Arc about this. You have factors of two of compute, where you're like, can the model do something like open up an account on AWS and make some money for itself? Some of the things that are obvious prerequisites to complete survival in the wild. Just set those thresholds very well below and then as you proceed upward from there, do kind of more and more rigorous tests and be more and more careful about what it is you're doing.

Dwarkesh Patel
On Constitution AI, who decides what the constitution for the next generation of models or a potentially superhuman model is? How is that actually written?

Dario Amodei
Initially to make the constitution, we just took some stuff that was broadly agreed on, like the UN declaration on Human Rights and some of the stuff from Apple's Terms of Service.

Stuff that's consensus on what's acceptable to say or what basic things are able to be included.

One, for future constitutions, we're looking into more participatory processes for making these. But beyond that, I don't think there should be one constitution for a model that everyone uses. The model's constitution should be very simple. It should only have very basic facts that everyone would agree on. Then there should be a lot of ways that you can customize, including appending constitutions. And beyond that, we're developing new methods. I'm not imagining that this or this alone is the method that we'll use to train superhuman AI. Many of the parts of capability training may be different, and so it could look very different.

There are levels above this. I'm pretty uncomfortable with: here's the AI's constitution, it's going to run the world. From just normal lessons from how societies work and how politics works, that strikes me as fanciful.

Even after we've mitigated the safety issues, any good future, even if it has all these security issues that we need to solve, it somehow needs to end with something that's more decentralized and less like a godlike super. I just don't think that ends well.

Dwarkesh Patel
What scientists from the Manhattan Project do you respect most in terms of, they acted most ethically under the constraints they were given. Is there one that comes to mind?

Dario Amodei
I don't know. There's a lot of answers you could give. I'm definitely a fan of Szilard for having kind of figured it out. He was then against the actual dropping of the bomb. I don't actually know the history well enough to have an opinion on whether the demonstration of the bomb could have ended the war. I mean that involves a bunch of facts about Imperial Japan that are complicated and that I'm not an expert on. But Szilard, he discovered this stuff early, he kept it secret, patented some of it and put it in the hands of the British Admiralty. He seemed to display the right kind of awareness as well as discovering stuff. It was when I read that book that when I wrote this big blob of compute doc and I only showed it to a few people and there were other docs that I showed to almost no one. I was a bit inspired by this.

Again, we could all get self aggrandizing here. Like we don't know if it's actually going to be something on par with the Manhattan project. This could all be just Silicon Valley people building technology and just having delusions of grandeur. I don't know how it's going to turn out.

Dwarkesh Patel
I mean, if the scaling stuff is true then it's bigger than the Manhattan Project.

Dario Amodei
Yeah, it certainly could be bigger. I think we should always maintain this attitude that it's really easy to fool yourself.

Dwarkesh Patel
If you're a physicist during World War II and you were asked by the government to contribute non replaceable research to the Manhattan Project, what do you think you would have said?

Dario Amodei
Given you're in a war with the Nazis, I don't really see much choice but to do it if it's possible. You have to figure it's going to be done within ten years or so by someone.

Dwarkesh Patel
Regarding cybersecurity, what should we make of the fact that there's a whole bunch of tech companies which have ordinary tech company security policy and it's not obvious that they've been hacked publicly. Coinbase still has its bitcoin. As far as I know my Gmail hasn't been leaked.

Should we take from that that current status quo tech company security practices are good enough for AGI or just simply that nobody has tried hard enough?

Dario Amodei
It would be hard for me to speak to current tech company practices and of course there may be many attacks that we don't know about, where things are stolen and then silently used. I think an indication of it is when someone really cares basically cares about attacking someone, then often the attacks happen.

Recently we saw that some fairly high officials of the US government had their email accounts hacked via Microsoft. Microsoft was providing the email accounts. Presumably that relayed information that was of great interest to foreign adversaries.

It seems to me at least that the evidence is more consistent with, when something is really high enough value, then someone acts and it's stolen. And my worry is that of course with AGI we'll get to a world where the value is seen as incredibly high. It'll be like stealing nuclear missiles or something. You can't be too careful on this stuff.

At every place that I've worked, I've pushed for cybersecurity to be better. One of my concerns about cybersecurity is, it's not something you can trumpet. A good dynamic with safety research is, you can get companies into a dynamic and I think we have, where you

can get them to compete to do the best safety research and use it as a recruiting point of competition or something. We used to do this all the time with interpretability and then sooner or later other orgs started recognizing the defect and started working on interpretability, whether or not that was a priority to them before.

But it's harder to do that with cybersecurity because a bunch of the stuff you have to do quietly. We did try to put out one post about it, but mostly you just see the results. A good norm would be people see these cybersecurity leaks from companies or leaks the model parameters or something and say they screwed up, that's bad. If I'm a safety person, I might not want to work there.

Of course, as soon as I say that, we'll probably have a security breach tomorrow. But that's part of the game here, that's part of trying to make things safe.

Dwarkesh Patel
I want to go back to the thing we're talking about earlier, where the ultimate level of cybersecurity required two to three years from now and whether it requires a bunker, are you actually expecting to be in a physical bunker in two to three years, or is that just a metaphor?

Dario Amodei
That's a metaphor. We're still figuring it out. Something I would think about is the security of the data center, which may not be in the same physical location as us, but we've worked very hard to make sure it's in the United States. But securing the physical data centers and the GPUs. If someone was really determined, some of the really expensive attacks just involve going into the data center and just trying to steal the data directly or as it's flowing from a data center to us. These data centers are going to have to be built in a very special way. Given the way things are scaling up, we're anyway heading to a world where the networks of data centers cost as much as aircraft carriers. They're already going to be pretty unusual objects but in addition to being unusual in terms of their ability to link together and train gigantic, gigantic models, they're also going to have to be very secure.

Dwarkesh Patel
Speaking of which, there's been rumors on the difficulty of procuring the power and the GPUs for the next generation of models. What has the process been like to secure the necessary components to do the next generation?

Dario Amodei
That's something I can't go into great detail about. I will say, people are thinking of industrial scale data centers and people are not thinking at the scale that these models are going to go to very soon. Whenever you do something at a scale where it's never been done before, every single component, every single thing has to be done in a new way than it was before.

And so you may run into problems with surprisingly simple components. Power is one that you mentioned.

Dwarkesh Patel
And is this something that Anthropic has to handle, or can you just outsource it?

Dario Amodei
For data centers, we work with cloud providers, for instance.

Dwarkesh Patel
What should we make about the fact that these models require so much training and the entire corpus of internet data in order to be subhuman?

Whereas GPT-4, there's been estimates that it was like 10^25 Flops or something, you can take these numbers with a grain of salt, but there's reports that the human brain, from the time it is born to the time a human being is 20 years old, is on the order of 10^14 Flops to simulate all those interactions.

We don't have to go into the particulars on those numbers, but should we be worried about how sample inefficient these models seem to be?

Dario Amodei
That's one of the remaining mysteries. One way you could phrase it is that the models are maybe two to three orders of magnitude smaller than the human brain. If you compare it to the number of synapses, while at the same time being trained on three to four more orders of magnitude of data. If you compare the number of words a human sees as they're developing to age 18, I don't remember exactly, but I think it's in the hundreds of millions, whereas for the models, we're talking about the hundreds of billions to the trillions. So what explains this? There are these offsetting things where the models are smaller, they need a lot more data. They're still below human level.

There's some way in which the analogy to the brain is not quite right or is breaking down or there's some missing factor. This is just like in physics, where we can't explain the Michelson-Morley experiment, or one of the other 19th century physics paradoxes. It's one thing we don't quite understand. Humans see so little data, and they still do fine.

One theory on it, it could be that it's like our other modalities. How do we get 10^14 bits into the human brain? Most of it is these images, and maybe a lot of what's going on inside the human brain is, our mental workspace involves all these simulated images or something like that.

But honestly, intellectually we have to admit that that's a weird thing that doesn't match up. And it's one reason I'm a bit skeptical of biological analogies. I thought in terms of them, like, five or six years ago, but now that we actually have these models in front of us as artifacts, it feels like almost all the evidence from that has been screened off by what we've seen. And what we've seen are models that are much smaller than the human brain and yet can do a lot of the things that humans can do, and yet, paradoxically, require a lot more data. Maybe we'll discover something that makes it all efficient, or maybe we'll understand why the discrepancy is present, but at the end of the day, I don't think it matters, right? If we keep scaling the way we are. I think what's more relevant at this point is just measuring the abilities of the model and seeing how far they are from humans, and they don't seem terribly far to me.

Dwarkesh Patel
Does this scaling picture and the big blob of compute more generally, underemphasize the role that algorithmic progress has played. When you composed the big blob of compute, you're presumably talking about LSTMs at that point, the scaling on that would not have you at Claude 2 at this point.

Are you underemphasizing the role that an improvement of the scale of Transformer could be having here, when you put it behind the label of scaling?

Dario Amodei
This big blob of compute document, which I still have not made public, I probably should for historical reasons. I don't think it would tell anyone anything they don't know now. But when I wrote it, I actually said, look, there are seven factors and I wasn't like, these are all the factors but just let me give some sense of the kinds of things that matter and what don't. There could be nine, there could be five. But the things I said were — Number of parameters matters. Scale of the model matters. Compute matters. Quantity of data matters. Quality of data matters. Loss function matters. Are you doing RL? Are you doing next word prediction? If your loss function isn't rich or doesn't incentivize the right thing, you won't get anything. Those were the key four ones, which I think are the core of the hypothesis.

But then I said three more things. One was symmetries, which is basically if your architecture doesn't take into account the right kinds of symmetries, it doesn't work or it's very inefficient. For example, convolutional neural networks take into account translational symmetry. LSTMs take into account time symmetry. But a weakness of LSTMs is that they can't attend over the whole context. So there's kind of this structural weakness. If a model isn't structurally capable of absorbing and managing things that happened in a far enough distant past, then it's like the compute doesn't flow. The spice doesn't flow. The blob has to be unencumbered. It's not going to work if you artificially close things off. And I think RNNs and LSTMs artificially close things off because they close you off to the distant past. Again, things need to flow freely. If they don't, it doesn't work.

And then I added a couple things. One of them was conditioning, which is if the thing you're optimizing with is just really numerically bad, you're going to have trouble. And so this is why atom works better than normal STD.

I'm forgetting what the 7th condition was, but it was similar to things like this, where if you set things up in a way that's set up to fail or that doesn't allow the compute to work in an uninhibited way, then it won't work. Transformers were kind of within that even though I can't remember if the transformer paper had been published, it was around the same time as I wrote that document. It might have been just before. It might have been just after.

Dwarkesh Patel
From that view it sounds like the way to think about these algorithmic progresses is not as increasing the power of the blob of compute, but simply getting rid of the artificial hindrances that older architectures have.

Dario Amodei
That's a little how I think about it. If you go back to Ilya's, the models want to learn, the compute wants to be free and it's being blocked in various ways where you don't understand that it's being blocked until you need to free it up.

Dwarkesh Patel
On that point, though, do you think that another thing on the scale of a transformer is coming down the pike to enable the next great iteration?

Dario Amodei
I think it's possible. People have worked on things like trying to model very long time dependencies or there's various different ideas where I could see that we're missing an efficient way of representing or dealing with something. I think those inventions are possible.

I guess my perspective would be, even if they don't happen, we're already on this very, very steep trajectory. Unless we're constantly trying to discover them, as are others, but things are already on such a fast trajectory, all that would do is speed up the trajectory even more, and probably not by that much because it's already going so fast.

Dwarkesh Patel
Is having an embodied version of a model at all important in terms of getting either data or progress?

Dario Amodei
I'd think of that less in terms of a new architecture and more in terms of a loss function like the data, the environments you're exposing yourself to end up being very different. That

could be important for learning some skills, although data acquisition is hard and so things have gone through the language route and I would guess will continue to go through the language route even as more is possible in terms of embodiment.

Dwarkesh Patel
And then the other possibilities you mentioned. RL, you can see it as...

Dario Amodei
We kind of already do RL with RLHF. Is this alignment? Is this capabilities? I always think in terms of the two snakes, they're often hard to distinguish. We already kind of use RL on these language models but I think we've used RL less in terms of getting them to take actions and do things in the world but when you take actions over a long period of time and understand the consequences of those actions only later, then RL is a typical tool we have for that. So I would guess that in terms of models taking action in the world, that RL will become a thing with all the power and all the safety issues that come with it.

Dwarkesh Patel
When you project out in the future, do you see the way in which these things will be integrated into productive supply chains? Do you see them talking with each other and criticizing each other and contributing to each other's output? Or is it just that one model one shots the answer or the work.

Dario Amodei
Models will undertake extended tasks. That will have to be the case. We may want to limit that to some extent because it may make some of the safety problems easier but some of that will be required.

In terms of our models talking to models or are they talking to humans? Again, this goes kind of out of the technical realm and into the sociocultural economic realm where my heuristic is always that it's very, very difficult to predict things. I feel like these scaling laws have been very predictable but then when you say like, when is there going to be a commercial explosion in these models? Or what's the form it's going to be? Or are the models going to do things instead of humans or pairing with humans? Certainly my track record on predicting these things is terrible but also looking around, I don't really see anyone whose track record is great.

Dwarkesh Patel
You mentioned how fast progress is happening, but also the difficulties of integrating within the existing economy into the way things work. Do you think there will be enough time to actually have large revenues from AI products before the next model is just so much better or we're in a different landscape entirely?

Dario Amodei
It depends what you mean by large. I think multiple companies are already in the 100 million to billion per year range. Will it get to the 100 billion or trillion range before? That stuff is just so hard to predict. And it's not even super well defined.

Right now there are companies that are throwing a lot of money at generative AI as customers. That's the right thing for them to do, and they'll find uses for it, but it doesn't mean they're finding uses or the best uses from day one. Even money changing hands is not quite the same thing as economic value being created.

Dwarkesh Patel
But surely you've thought about this from the perspective of Anthropic, where if these things are happening so fast, then it should be an insane valuation, right?

Dario Amodei
Even us who have not been super focused on commercialization and more on safety, the graph goes up and it goes up relatively quickly. I can only imagine what's happening at the orgs where this is their singular focus. It's certainly happening fast but it's an exponential from the small base while the technology itself is moving fast.

It's a race between how fast the technology is getting better and how fast it's integrated into the economy. And I think that's just a very unstable and turbulent process. Both things are going to happen fast but if you ask me exactly how it's going to play out, exactly what order things are going to happen, I don't know. And I'm skeptical of the ability to predict.

Dwarkesh Patel
I'm curious. With regards to Anthropic specifically, you're a public benefit corporation and rightfully so, you want to make sure that this is an important technology. Obviously, the only thing you want to care about is not shareholder value.

But how do you talk to investors who are putting in hundreds of millions, billions of dollars of money? How do you get them to put in this amount of money without the shareholder value being the main concern?

Dario Amodei
I think the LTBT (Long Term Benefit Trust) is the right thing on this. We're going to talk more about the LTBT, but some version of that has been in development since the beginning of Anthropic, even formally. Even as the body has changed, from the beginning, it was like, this body is going to exist and it's unusual.

Every traditional investor who invests in Anthropic looks at this. Some of them are just like, whatever, you run your company how you want. Some of them are like, oh my god, this body

of random people could move Anthropic in a direction that's totally contrary to shareholder value. Now there are legal limits on that, of course, but we have to have this conversation with every investor. And then it gets into a conversation of, well, what are the kinds of things that we might do that would be contrary to the interests of traditional investors. And just having those conversations has helped get everyone on the same page.

Dwarkesh Patel
I want to talk about the fact that so many of the founders and the employees at Anthropic are physicists. We talked in the beginning about the scaling laws and how the power laws from physics are something you see here, but what are the actual approaches and ways of thinking from physics that seem to have carried over so well? Is that notion of effective theory super useful? What is going on here?

Dario Amodei
Part of it is just that physicists learn things really fast. We have generally found that if we hire someone who is a Physics PhD or something, that they can learn ML and contribute just very quickly in most cases. And because several of our founders myself, Jared Kaplan, Sam McCandlish were physicists, we knew a lot of other physicists, and so we were able to hire them. And now there might be 30 or 40 of them here. ML is not still not yet a field that has an enormous amount of depth, and so they've been able to get up to speed very quickly.

Dwarkesh Patel
Are you concerned that there's a lot of people who would have been doing physics or something, they would've gone into finance instead and since Anthropic exists, they have now been recruited to go into AI. You obviously care about AI safety, but maybe in the future they leave and they get funded to do their own thing. Is that a concern that you're bringing more people into the ecosystem here?

Dario Amodei
There's a broad set of actions, like we're causing GPUs to exist. There's a lot of side effects that you can't currently control or that you just incur if you buy into the idea that you need to build frontier models. And that's one of them. A lot of them would have happened anyway. I mean, finance was a hot thing 20 years ago, so physicists were doing it. Now ML is a hot thing, and it's not like we've caused them to do it when they had no interest previously. But again, at the margin, you're bidding things up, and a lot of that would have happened anyway. Some of it wouldn't but it's all part of the calculus.

Dwarkesh Patel
Do you think that Claude has conscious experience? How likely do you think that is?

Dario Amodei
This is another of these questions that just seems very unsettled and uncertain. One thing I'll tell you is I used to think that we didn't have to worry about this at all until models were operating in rich environments, like not necessarily embodied, but they needed to have a reward function and have a long lived experience. I still think that might be the case, but the more we've looked at these language models and particularly looked inside them to see things like induction heads, a lot of the cognitive machinery that you would need for active agents already seems present in the base language models. So I'm not quite as sure as I was before that we're missing enough of the things that you would need. I think today's models just probably aren't smart enough that we should worry about this too much but I'm not 100% sure about this, and I do think in a year or two, this might be a very real concern.

Dwarkesh Patel
What would change if you found out that they are conscious? Are you worried that you're pushing the negative gradient to suffering?

Dario Amodei
Conscious, again, is one of these words that I suspect will not end up having a well defined.. I suspect that's a spectrum. Let's say we discover that I should care about Claude's experience as much as I should care about a dog or a monkey or something. I would be kind of worried.

I don't know if their experience is positive or negative. Unsettlingly I also don't know I wouldn't know if any intervention that we made was more likely to make Claude have a positive versus negative experience versus not having one.

If there's an area that is helpful with this, it's maybe mechanistic interpretability because I think of it as neuroscience for models. It's possible that we could shed some light on this. Although it's not a straightforward factual question. It depends what we mean and what we value.

Dwarkesh Patel
We talked about this initially, but I want to get more specific. We talked initially about now that you're seeing these capabilities ramp up within the human spectrum, you think that the human spectrum is wider than we thought but more specifically, how is the way you think about human intelligence different. The way you're seeing these marginally useful abilities emerge? How does that change your picture of what intelligence is?

Dario Amodei
For me, the big realization on what intelligence is came with the blob of compute thing. There might be all these separate modules. There might be all this complexity. Rich Sutton called it The Bitter Lesson. It has many names. It's been called the scaling hypothesis. The

first few people who figured it out was around 2017. You could go further back. I think Shane Legg was maybe the first person who really knew it, maybe Ray Kurzweil, although in a very vague way. But the number of people who understood it went up a lot around 2014 to 2017.

I think that was the big realization. How did intelligence evolve? If you don't need very specific conditions to create it, if you can create it just from the right kind of gradient and loss signal, then of course it's not so mysterious how it all happened. It had this click of scientific understanding.

In terms of watching what the models can do, how has it changed my view of human intelligence? I wish I had something more intelligent to say on that. One thing that's been surprising is I thought things might click into place a little more than they do. I thought different cognitive abilities might all be connected and there was more of one secret behind them. But the model just learns various things at different times. It can be very good at coding but it can't quite prove the prime number theorem yet. And I guess it's a little bit the same for humans, although it's weird the juxtaposition of things it can do and not. I guess the main lesson is having theories of intelligence or how intelligence works. A lot of these words just dissolve into a continuum. They just kind of dematerialize. I think less in terms of intelligence and more in terms of what we see in front of us.

Dwarkesh Patel
Two things are really surprising to me. One is how discrete these different paths of intelligent things that contribute to loss are rather than just being one reasoning circuit or one general intelligence. And the other surprising and interesting thing is, many years from now, it'll be one of those things that you'll wonder why it wasn't obvious to you? If you're seeing these smooth scaling curves, why were you not completely convinced at the time?

You've been less public than the CEOs of other AI companies. You're not posting on Twitter, you're not doing a lot of podcasts except for this one. What gives? Why are you off the radar?

Dario Amodei
I aspire to this and I'm proud of this. If people think of me as boring and low profile, this is actually kind of what I want. I've just seen cases with a number of people I've worked with, where attaching your incentives very strongly to the approval or cheering of a crowd can destroy your mind, and in some cases, it can destroy your soul.

I've deliberately tried to be a little bit low profile because I want to defend my ability to think about things intellectually in a way that's different from other people and isn't tinged by the approval of other people. I've seen cases of folks who are deep learning skeptics, and they become known as deep learning skeptics on Twitter. And then even as it starts to become

clear to me, they've sort of changed their mind. This is their thing on Twitter, and they can't change their Twitter persona and so forth and so on.

I don't really like the trend of personalizing companies. The whole cage match between CEOs approach. I think it distracts people from the actual merits and concerns of the company in question. I want people to think in terms of the nameless, bureaucratic institution and its incentives more than they think in terms of me. Everyone wants a friendly face, but actually, friendly faces can be misleading.

Dwarkesh Patel
Okay, well, in this case, this will be a misleading interview because this has been a lot of fun.

Dario Amodei
Indeed.

Dwarkesh Patel
Yeah, this has been a blast. I'm super glad you came on the podcast and hope people enjoyed it.

Dario Amodei
Thanks for having me.