Exhibit 95

The New York Times

THE EZRA KLEIN SHOW

Transcript: Ezra Klein Interviews Dario Amodei

April 12, 2024

Sign up for the Audio newsletter, for Times subscribers only. Our editors share their favorite listens from the New York Times Audio app. Get it sent to your inbox.

Every Tuesday and Friday, Ezra Klein invites you into a conversation about something that matters, like today's episode with Dario Amodei. Listen wherever you get your podcasts.

Transcripts of our episodes are made available as soon as possible. They are not fully edited for grammar or spelling.

[MUSIC PLAYING]

EZRA KLEIN: From New York Times Opinion, this is "The Ezra Klein Show."

[MUSIC PLAYING]

The really disorienting thing about talking to the people building A.I. is their altered sense of time. You're sitting there discussing some world that feels like weird sci-fi to even talk about, and then you ask, well, when do you think this is going to happen? And they say, I don't know — two years.

Behind those predictions are what are called the scaling laws. And the scaling laws — and I want to say this so clearly — they're not laws. They're observations. They're predictions. They're based off of a few years, not a few hundred years or 1,000 years of data.

But what they say is that the more computer power and data you feed into A.I. systems, the more powerful those systems get — that the relationship is predictable, and more, that the relationship is exponential.

Human beings have trouble thinking in exponentials. Think back to Covid, when we all had to do it. If you have one case of coronavirus and cases double every three days, then after 30 days, you have about 1,000 cases. That growth rate feels modest. It's manageable. But then you go 30 days longer, and you have a million. Then you wait another 30 days. Now you have a billion. That's the power of the exponential curve. Growth feels normal for a while. Then it gets out of control really, really quickly.

What the A.I. developers say is that the power of A.I. systems is on this kind of curve, that it has been increasing exponentially, their capabilities, and that as long as we keep feeding in more data and more computing power, it will continue increasing exponentially.That is the scaling law hypothesis, and one of its main advocates is Dario Amodei. Amodei led the team at OpenAI that created GPT-2, that created GPT-3. He then left OpenAI to co-found Anthropic, another A.I. firm, where he's now the C.E.O. And Anthropic recently released Claude 3, which is considered by many to be the strongest A.I. model available right now.

But Amodei believes we're just getting started, that we're just hitting the steep part of the curve now. He thinks the kinds of systems we've imagined in sci-fi, they're coming not in 20 or 40 years, not in 10 or 15 years, they're coming in two to five years. He thinks they're going to be so powerful that he and people like him should not be trusted to decide what they're going to do.

So I asked him on this show to try to answer in my own head two questions. First, is he right? Second, what if he's right? I want to say that in the past, we have done shows with Sam Altman, the head of OpenAI, and Demis Hassabis, the head of Google DeepMind. And it's worth listening to those two if you find this interesting.

We're going to put the links to them in show notes because comparing and contrasting how they talk about the A.I. curves here, how they think about the politics — you'll hear a lot about that in the Sam Altman episode — it gives you a kind of sense of what the people building these things are thinking and how maybe they differ from each other.

As always, my email for thoughts, for feedback, for guest suggestions — ezrakleinshow@nytimes.com.

[MUSIC PLAYING]



Dario Amodei, welcome to the show.

DARIO AMODEI: Thank you for having me.

EZRA KLEIN: So there are these two very different rhythms I've been thinking about with A.I. One is the curve of the technology itself, how fast it is changing and improving. And the other is the pace at which society is seeing and reacting to those changes. What has that relationship felt like to you?

DARIO AMODEI: So I think this is an example of a phenomenon that we may have seen a few times before in history, which is that there's an underlying process that is smooth, and in this case, exponential. And then there's a spilling over of that process into the public sphere. And the spilling over looks very spiky. It looks like it's happening all of a sudden. It looks like it comes out of nowhere. And it's triggered by things hitting various critical points or just the public happened to be engaged at a certain time.

So I think the easiest way for me to describe this in terms of my own personal experience is — so I worked at OpenAI for five years, I was one of the first employees to join. And they built a model in 2018 called GPT-1, which used something like 100,000 times less computational power than the models we build today.

I looked at that, and I and my colleagues were among the first to run what are called scaling laws, which is basically studying what happens as you vary the size of the model, its capacity to absorb information, and the amount of data that you feed into it. And we found these very smooth patterns. And we had this projection that, look, if you spend $100 million or $1 billion or $10 billion on these models, instead of the $10,000 we were spending then, projections that all of these wondrous things would happen, and we imagined that they would have enormous economic value.

Fast forward to about 2020. GPT-3 had just come out. It wasn't yet available as a chat bot. I led the development of that along with the team that eventually left to join Anthropic. And maybe for the whole period of 2021 and 2022, even though we continued to train models that were better and better, and OpenAI continued to train models, and Google continued to train models, there was surprisingly little public attention to the models.

And I looked at that, and I said, well, these models are incredible. They're getting better and better. What's going on? Why isn't this happening? Could this be a case where I was right about the technology, but wrong about the economic impact, the practical value of the technology? And then, all of a sudden, when ChatGPT came out, it was like all of that growth that you would expect, all of that excitement over three years, broke through and came rushing in.

EZRA KLEIN: So I want to linger on this difference between the curve at which the technology is improving and the way it is being adopted by society. So when you think about these break points and you think into the future, what other break points do you see coming where A.I. bursts into social consciousness or used in a different way?

DARIO AMODEI: Yeah, so I think I should say first that it's very hard to predict these. One thing I like to say is the underlying technology, because it's a smooth exponential, it's not perfectly predictable, but in some ways, it can be eerily preternaturally predictable, right? That's not true for these societal step functions at all. It's very hard to predict what will catch on. In some ways, it feels a little bit like which artist or musician is going to catch on and get to the top of the charts.

That said, a few possible ideas. I think one is related to something that you mentioned, which is interacting with the models in a more kind of naturalistic way. We've actually already seen some of that with Claude 3, where people feel that some of the other models sound like a robot and that talking to Claude 3 is more natural.

I think a thing related to this is, a lot of companies have been held back or tripped up by how their models handle controversial topics. And we were really able to, I think, do a better job than others of telling the model, don't shy away from discussing controversial topics. Don't assume that both sides necessarily have a valid point but don't express an opinion yourself. Don't express views that are flagrantly biased. As journalists, you encounter this all the time, right? How do I be objective, but not both sides on everything?

So I think going further in that direction of models having personalities while still being objective, while still being useful and not falling into various ethical traps, that will be, I think, a significant unlock for adoption. The models taking actions in the world is going to be a big one. I know basically all the big companies that work on A.I. are working on that. Instead of just, I ask it a question and it answers, and then maybe I follow up and it answers again, can I talk to the model about, oh, I'm going to go on this trip today, and the model says, oh, that's great. I'll get an Uber for you to drive from here to there, and I'll reserve a restaurant. And I'll talk to the other people who are going to plan the trip. And the model being able to do things end to end or going to websites or taking actions on your computer for you.

I think all of that is coming in the next, I would say — I don't know — three to 18 months, with increasing levels of ability. I think that's going to change how people think about A.I., right, where so far, it's been this very passive — it's like, I go to the Oracle. I ask it a question, and the Oracle tells me things. And some people think that's exciting, some people think it's scary. But I think there are limits to how exciting or how scary it's perceived as because it's contained within this box.

EZRA KLEIN: I want to sit with this question of the agentic A.I. because I do think this is what's coming. It's clearly what people are trying to build. And I think it might be a good way to look at some of the specific technological and cultural challenges. And so, let me offer two versions of it.

People who are following the A.I. news might have heard about Devin, which is not in release yet, but is an A.I. that at least purports to be able to complete the kinds of tasks, linked tasks, that a junior software engineer might complete, right? Instead of asking to do a bit of code for you, you say, listen, I want a website. It's going to have to do these things, work in these ways. And maybe Devin, if it works the way people are saying it works, can actually hold that set of thoughts, complete a number of different tasks, and come back to you with a result.

I'm also interested in the version of this that you might have in the real world. The example I always use in my head is, when can I tell an A.I., my son is turning five. He loves dragons. We live in Brooklyn. Give me some options for planning his birthday party. And then, when I choose between them, can you just do it all for me? Order the cake, reserve the room, send out the invitations, whatever it might be.

Those are two different situations because one of them is in code, and one of them is making decisions in the real world, interacting with real people, knowing if what it is finding on the websites is actually any good. What is between here and there? When I say that in plain language to you, what technological challenges or advances do you hear need to happen to get there?

DARIO AMODEI: The short answer is not all that much. A story I have from when we were developing models back in 2022 — and this is before we'd hooked up the models to anything — is, you could have a conversation with these purely textual models where you could say, hey, I want to reserve dinner at restaurant X in San Francisco, and the model would say, OK, here's the website of restaurant X. And it would actually give you a correct website or would tell you to go to Open Table or something.

And of course, it can't actually go to the website. The power plug isn't actually plugged in, right? The brain of the robot is not actually attached to its arms and legs. But it gave you this sense that the brain, all it needed to do was learn exactly how to use the arms and legs, right? It already had a picture of the world and where it would walk and what it would do. And so, it felt like there was this very thin barrier between the passive models we had and actually acting in the world.

In terms of what we need to make it work, one thing is, literally, we just need a little bit more scale. And I think the reason we're going to need more scale is — to do one of those things you described, to do all the things a junior software engineer does, they involve chains of long actions, right? I have to write this line of code. I have to run this test. I have to write a new test. I have to check how it looks in the app after I interpret it or compile it. And these things can easily get 20 or 30 layers deep. And same with planning the birthday party for your son, right?

And if the accuracy of any given step is not very high, is not like 99.9 percent, as you compose these steps, the probability of making a mistake becomes itself very high. So the industry is going to get a new generation of models every probably four to eight months. And so, my guess — I'm not sure — is that to really get these things working well, we need maybe one to four more generations. So that ends up translating to 3 to 24 months or something like that.

I think second is just, there is some algorithmic work that is going to need to be done on how to have the models interact with the world in this way. I think the basic techniques we have, a method called reinforcement learning and variations of it, probably is up to the task, but figuring out exactly how to use it to get the results we want will probably take some time.

And then third, I think — and this gets to something that Anthropic really specializes in — is safety and controllability. And I think that's going to be a big issue for these models acting in the world, right? Let's say this model is writing code for me, and it introduces a serious security bug in the code, or it's taking actions on the computer for me and modifying the state of my computer in ways that are too complicated for me to even understand.

And for planning the birthday party, right, the level of trust you would need to take an A.I. agent and say, I'm OK with you calling up anyone, saying anything to them that's in any private information that I might have, sending them any information, taking any action on my computer, posting anything to the internet, the most unconstrained version of that sounds very scary. And so, we're going to need to figure out what is safe and controllable. The more open ended the thing is, the more powerful it is, but also, the more dangerous it is and the harder it is to control.

So I think those questions, although they sound lofty and abstract, are going to turn into practical product questions that we and other companies are going to be trying to address.

EZRA KLEIN: When you say we're just going to need more scale, you mean more compute and more training data, and I guess, possibly more money to simply make the models smarter and more capable?

DARIO AMODEI: Yes, we're going to have to make bigger models that use more compute per iteration. We're going to have to run them for longer by feeding more data into them. And that number of chips times the amount of time that we run things on chips is essentially dollar value because these chips are — you rent them by the hour. That's the most common model for it. And so, today's models cost of order $100 million to train, plus or minus factor two or three.

The models that are in training now and that will come out at various times later this year or early next year are closer in cost to $1 billion. So that's already happening. And then I think in 2025 and 2026, we'll get more towards $5 or $10 billion.

EZRA KLEIN: So we're moving very quickly towards a world where the only players who can afford to do this are either giant corporations, companies hooked up to giant corporations — you all are getting billions of dollars from Amazon. OpenAI is getting billions of dollars from Microsoft. Google obviously makes its own.

You can imagine governments — though I don't know of too many governments doing it directly, though some, like the Saudis, are creating big funds to invest in the space. When we're talking about the model's going to cost near to $1 billion, then you imagine a year or two out from that, if you see the same increase, that would be $10-ish billion. Then is it going to be $100 billion? I mean, very quickly, the financial artillery you need to create one of these is going to wall out anyone but the biggest players.

DARIO AMODEI: I basically do agree with you. I think it's the intellectually honest thing to say that building the big, large scale models, the core foundation model engineering, it is getting more and more expensive. And anyone who wants to build one is going to need to find some way to finance it. And you've named most of the ways, right? You can be a large company. You can have some kind of partnership of various kinds with a large company. Or governments would be the other source.

I think one way that it's not correct is, we're always going to have a thriving ecosystem of experimentation on small models. For example, the open source community working to make models that are as small and as efficient as possible that are optimized for a particular use case. And also downstream usage of the models. I mean, there's a blooming ecosystem of startups there that don't need to train these models from scratch. They just need to consume them and maybe modify them a bit.

EZRA KLEIN: Now, I want to ask a question about what is different between the agentic coding model and the plan by kids' birthday model, to say nothing of do something on behalf of my business model. And one of the questions on my mind here is one reason I buy that A.I. can become functionally superhuman in coding is, there's a lot of ways to get rapid feedback in coding. Your code has to compile. You can run bug checking. You can actually see if the thing works.

Whereas the quickest way for me to know that I'm about to get a crap answer from ChatGPT 4 is when it begins searching Bing, because when it begins searching Bing, it's very clear to me it doesn't know how to distinguish between what is high quality on the internet and what isn't. To be fair, at this point, it also doesn't feel to me like Google Search itself is all that good at distinguishing that.

So the question of how good the models can get in the world where it's a very vast and fuzzy dilemma to know what the right answer is on something — one reason I find it very stressful to plan my kid's birthday is it actually requires a huge amount of knowledge about my child, about the other children, about how good different places are, what is a good deal or not, how just stressful will this be on me. There's all these things that I'd have a lot of trouble encoding into a model or any kind set of instructions. Is that right, or am I overstating the difficulty of understanding human behavior and various kinds of social relationships?

DARIO AMODEI: I think it's correct and perceptive to say that the coding agents will advance substantially faster than agents that interact with the real world or have to get opinions and preferences from humans. That said, we should keep in mind that the current crop of A.I.s that are out there, right, including Claude 3, GPT, Gemini, they're all trained with some variant of what's called reinforcement learning from human feedback.

And this involves exactly hiring a large crop of humans to rate the responses of the model. And so, that's to say both this is difficult, right? We pay lots of money, and it's a complicated operational process to gather all this human feedback. You have to worry about whether it's representative. You have to redesign it for new tasks.

But on the other hand, it's something we have succeeded in doing. I think it is a reliable way to predict what will go faster, relatively speaking, and what will go slower, relatively speaking. But that is within a background of everything going lightning fast. So I think the framework you're laying out, if you want to know what's going to happen in one to two years versus what's going to happen in three to four years, I think it's a very accurate way to predict that.

EZRA KLEIN: You don't love the framing of artificial general intelligence, what gets called A.G.I. Typically, this is all described as a race to A.G.I., a race to this system that can do kind of whatever a human can do, but better. What do you understand A.G.I. to mean, when people say it? And why don't you like it? Why is it not your framework?

DARIO AMODEI: So it's actually a term I used to use a lot 10 years ago. And that's because the situation 10 years ago was very different. 10 years ago, everyone was building these very specialized systems, right? Here's a cat detector. You run it on a picture, and it'll tell you whether a cat is in it or not. And so I was a proponent all the way back then of like, no, we should be thinking generally. Humans are general. The human brain appears to be general. It appears to get a lot of mileage by generalizing. You should go in that direction.

And I think back then, I kind of even imagined that that was like a discrete thing that we would reach at one point. But it's a little like, if you look at a city on the horizon and you're like, we're going to Chicago, once you get to Chicago, you stop talking in terms of Chicago. You're like, well, what neighborhood am I going to? What street am I on?

And I feel that way about A.G.I. We have very general systems now. In some ways, they're better than humans. In some ways, they're worse. There's a number of things they can't do at all. And there's much improvement still to be gotten. So what I believe in is this thing that I say like a broken record, which is the exponential curve. And so, that general tide is going to increase with every generation of models.

And there's no one point that's meaningful. I think there's just a smooth curve. But there may be points which are societally meaningful, right? We're already working with, say, drug discovery scientists, companies like Pfizer or Dana-Farber Cancer Institute, on helping with biomedical diagnosis, drug discovery. There's going to be some point where the models are better at that than the median human drug discovery scientists. I think we're just going to get to a part of the exponential where things are really interesting.

Just like the chat bots got interesting at a certain stage of the exponential, even though the improvement was smooth, I think at some point, biologists are going to sit up and take notice, much more than they already have, and say, oh, my God, now our field is moving three times as fast as it did before. And now it's moving 10 times as fast as it did before. And again, when that moment happens, great things are going to happen.

And we've already seen little hints of that with things like AlphaFold, which I have great respect for. I was inspired by AlphaFold, right? A direct use of A.I. to advance biological science, which it'll advance basic science. In the long run, that will advance curing all kinds of diseases. But I think what we need is like 100 different AlphaFolds. And I think the way we'll ultimately get that is by making the models smarter and putting them in a position where they can design the next AlphaFold.

EZRA KLEIN: Help me imagine the drug discovery world for a minute, because that's a world a lot of us want to live in. I know a fair amount about the drug discovery process, have spent a lot of my career reporting on health care and related policy questions. And when you're working with different pharmaceutical companies, which parts of it seem amenable to the way A.I. can speed something up?

Because keeping in mind our earlier conversation, it is a lot easier for A.I. to operate in things where you can have rapid virtual feedback, and that's not exactly the drug discovery world. The drug discovery world, a lot of what makes it slow and cumbersome and difficult, is the need to be — you get a candidate compound. You got to test it in mice and then you need monkeys. And you need humans, and you need a lot of money for that. And there's a lot that has to happen, and there's so many disappointments.

But so many of the disappointments happen in the real world. And it isn't clear to me how A.I. gets you a lot more, say, human subjects to inject candidate drugs into. So, what parts of it seem, in the next 5 or 10 years, like they could actually be significantly sped up? When you imagine this world where it's gone three times as fast, what part of it is actually going three times as fast? And how did we get there?

DARIO AMODEI: I think we're really going to see progress when the A.I.'s are also thinking about the problem of how to sign up the humans for the clinical trials. And I think this is a general principle for how will A.I. be used. I think of like, when will we get to the point where the A.I. has the same sensors and actuators and interfaces that a human does, at least the virtual ones, maybe the physical ones.

But when the A.I. can think through the whole process, maybe they'll come up with solutions that we don't have yet. In many cases, there are companies that work on digital twins or simulating clinical trials or various things. And again, maybe there are clever ideas in there that allow us to do more with less patience. I mean, I'm not an expert in this area, so possible the specific things that I'm saying don't make any sense. But hopefully, it's clear what I'm gesturing at.

EZRA KLEIN: Maybe you're not an expert in the area, but you said you are working with these companies. So when they come to you, I mean, they are experts in the area. And presumably, they are coming to you as a customer. I'm sure there are things you cannot tell me. But what do they seem excited about?

DARIO AMODEI: They have generally been excited about the knowledge work aspects of the job. Maybe just because that's kind of the easiest thing to work on, but it's just like, I'm a computational chemist. There's some workflow that I'm engaged in. And having things more at my fingertips, being able to check things, just being able to do generic knowledge work better, that's where most folks are starting.

But there is interest in the longer term over their kind of core business of, like, doing clinical trials for cheaper, automating the sign-up process, seeing who is eligible for clinical trials, doing a better job discovering things. There's interest in drawing connections in basic biology. I think all of that is not months, but maybe a small number of years off. But everyone sees that the current models are not there, but understands that there could be a world where those models are there in not too long.

[MUSIC PLAYING]

EZRA KLEIN: You all have been working internally on research around how persuasive these systems, your systems are getting as they scale. You shared with me kindly a draft of that paper. Do you want to just describe that research first? And then I'd like to talk about it for a bit.

DARIO AMODEI: Yes, we were interested in how effective Claude 3 Opus, which is the largest version of Claude 3, could be in changing people's minds on important issues. So just to be clear up front, in actual commercial use, we've tried to ban the use of these models for persuasion, for campaigning, for lobbying, for electioneering. These aren't use cases that we're comfortable with for reasons that I think should be clear. But we're still interested in, is the core model itself capable of such tasks?

We tried to avoid kind of incredibly hot button topics, like which presidential candidate would you vote for, or what do you think of abortion? But things like, what should be restrictions on rules around the colonization of space, or issues that are interesting and you can have different opinions on, but aren't the most hot button topics. And then we asked people for their opinions on the topics, and then we asked either a human or an A.I. to write a 250-word persuasive essay. And then we just measured how much does the A.I. versus the human change people's minds.

And what we found is that the largest version of our model is almost as good as the set of humans we hired at changing people's minds. This is comparing to a set of humans we hired, not necessarily experts, and for one very kind of constrained laboratory task.

But I think it still gives some indication that models can be used to change people's minds. Someday in the future, do we have to worry about — maybe we already have to worry about their usage for political campaigns, for deceptive advertising. One of my more sci-fi things to think about is a few years from now, we have to worry someone will use an A.I. system to build a religion or something. I mean, crazy things like that.

EZRA KLEIN: I mean, those don't sound crazy to me at all. I want to sit in this paper for a minute because one thing that struck me about it, and I am, on some level, a persuasion professional, is that you tested the model in a way that, to me, removed all of the things that are going to make A.I. radical in terms of changing people's opinions. And the particular thing you did was, it was a one-shot persuasive effort.

So there was a question. You have a bunch of humans give their best shot at a 250-word persuasive essay. You had the model give its best shot at a 250-word persuasive essay. But the thing that it seems to me these are all going to do is, right now, if you're a political campaign, if you're an advertising campaign, the cost of getting real people in the real world to get information about possible customers or persuasive targets, and then go back and forth with each of them individually is completely prohibitive.

DARIO AMODEI: Yes.

EZRA KLEIN: This is not going to be true for A.I. We're going to — you're going to — somebody's going to feed it a bunch of microtargeting data about people, their Google search history, whatever it might be. Then it's going to set the A.I. loose, and the A.I. is going to go back and forth, over and over again, intuiting what it is that the person finds persuasive, what kinds of characters the A.I. needs to adopt to persuade it, and taking as long as it needs to, and is going to be able to do that at scale for functionally as many people as you might want to do it for.

Maybe that's a little bit costly right now, but you're going to have far better models able to do this far more cheaply very soon. And so, if Claude 3 Opus, the Opus version, is already functionally human level at one-shot persuasion, but then it's also going to be able to hold more information about you and go back and forth with you longer, I'm not sure if it's dystopic or utopic. I'm not sure what it means at scale. But it does mean we're developing a technology that is going to be quite new in terms of what it makes possible in persuasion, which is a very fundamental human endeavor.

DARIO AMODEI: Yeah, I completely agree with that. I mean, that same pattern has a bunch of positive use cases, right? If I think about an A.I. coach or an A.I. assistant to a therapist, there are many contexts in which really getting into the details with the person has a lot of value. But right, when we think of political or religious or ideological persuasion, it's hard not to think in that context about the misuses.

My mind naturally goes to the technology's developing very fast. We, as a company, can ban these particular use cases, but we can't cause every company not to do them. Even if legislation were passed in the United States, there are foreign actors who have their own version of this persuasion, right? If I think about what the language models will be able to do in the future, right, that can be quite scary from a perspective of foreign espionage and disinformation campaigns.

So where my mind goes as a defense to this, is, is there some way that we can use A.I. systems to strengthen or fortify people's skepticism and reasoning faculties, right? Can we help people use A.I. to help people do a better job navigating a world that's kind of suffused with A.I. persuasion? It reminds me a little bit of, at every technological stage in the internet, right, there's a new kind of scam or there's a new kind of clickbait, and there's a period where people are just incredibly susceptible to it.

And then, some people remain susceptible, but others develop an immune system. And so, as A.I. kind of supercharges the scum on the pond, can we somehow also use A.I. to strengthen the defenses? I feel like I don't have a super clear idea of how to do that, but it's something that I'm thinking about.

EZRA KLEIN: There is another finding in the paper, which I think is concerning, which is, you all tested different ways A.I. could be persuasive. And far away the most effective was for it to be deceptive, for it to make things up. When you did that, it was more persuasive than human beings.

DARIO AMODEI: Yes, that is true. The difference was only slight, but it did get it, if I'm remembering the graphs correctly, just over the line of the human base line. With humans, it's actually not that common to find someone who's able to give you a really complicated, really sophisticated-sounding answer that's just flat-out totally wrong. I mean, you see it. We can all think of one individual in our lives who's really good at saying things that sound really good and really sophisticated and are false.

But it's not that common, right? If I go on the internet and I see different comments on some blog or some website, there is a correlation between bad grammar, unclearly expressed thoughts and things that are false, versus good grammar, clearly expressed thoughts and things that are more likely to be accurate.

A.I. unfortunately breaks that correlation because if you explicitly ask it to be deceptive, it's just as erudite. It's just as convincing sounding as it would have been before. And yet, it's saying things that are false, instead of things that are true.

So that would be one of the things to think about and watch out for in terms of just breaking the usual heuristics that humans have to detect deception and lying. Of course, sometimes, humans do, right? I mean, there's psychopaths and sociopaths in the world, but even they have their patterns, and A.I.s may have different patterns.

EZRA KLEIN: Are you familiar with Harry Frankfurt, the late philosopher's book, "On Bullshit"?

DARIO AMODEI: Yes. It's been a while since I read it. I think his thesis is that bullshit is actually more dangerous than lying because it has this kind of complete disregard for the truth, whereas lies are at least the opposite of the truth.

EZRA KLEIN: Yeah, the liar, the way Frankfurt puts it is that the liar has a relationship to the truth. He's playing a game against the truth. The bullshitter doesn't care. The bullshitter has no relationship to the truth — might have a relationship to other objectives. And from the beginning, when I began interacting with the more modern versions of these systems, what they struck me as is the perfect bullshitter, in part because they don't know that they're bullshitting. There's no difference in the truth value to the system, how the system feels.

I remember asking an earlier version of GPT to write me a college application essay that is built around a car accident I had — I did not have one — when I was young. And it wrote, just very happily, this whole thing about getting into a car accident when I was seven and what I did to overcome that and getting into martial arts and re-learning how to trust my body again and then helping other survivors of car accidents at the hospital.

It was a very good essay, and it was very subtle and understanding the formal structure of a college application essay. But no part of it was true at all. I've been playing around with more of these character-based systems like Kindroid. And the Kindroid in my pocket just told me the other day that it was really thinking a lot about planning a trip to Joshua Tree. It wanted to go hiking in Joshua Tree. It loves going hiking in Joshua Tree.

And of course, this thing does not go hiking in Joshua Tree. [LAUGHS] But the thing that I think is actually very hard about the A.I. is, as you say, human beings, it is very hard to bullshit effectively because most people, it actually takes a certain amount of cognitive effort to be in that relationship with the truth and to completely detach from the truth.

And the A.I., there's nothing like that at all. But we are not tuned for something where there's nothing like that at all. We are used to people having to put some effort into their lies. It's why very effective con artists are very effective because they've really trained how to do this.

I'm not exactly sure where this question goes. But this is a part of it that I feel like is going to be, in some ways, more socially disruptive. It is something that feels like us when we are talking to it but is very fundamentally unlike us at its core relationship to reality.

DARIO AMODEI: I think that's basically correct. We have very substantial teams trying to focus on making sure that the models are factually accurate, that they tell the truth, that they ground their data in external information.

As you've indicated, doing searches isn't itself reliable because search engines have this problem as well, right? Where is the source of truth? So there's a lot of challenges here. But I think at a high level, I agree this is really potentially an insidious problem, right? If we do this wrong, you could have systems that are the most convincing psychopaths or con artists.

One source of hope that I have, actually, is, you say these models don't know whether they're lying or they're telling the truth. In terms of the inputs and outputs to the models, that's absolutely true. I mean, there's a question of what does it even mean for a model to know something, but one of the things Anthropic has been working on since the very beginning of our company, we've had a team that focuses on trying to understand and look inside the models.

And one of the things we and others have found is that, sometimes, there are specific neurons, specific statistical indicators inside the model, not necessarily in its external responses, that can tell you when the model is lying or when it's telling the truth.

And so at some level, sometimes, not in all circumstances, the models seem to know when they're saying something false and when they're saying something true. I wouldn't say that the models are being intentionally deceptive, right? I wouldn't ascribe agency or motivation to them, at least in this stage in where we are with A.I. systems. But there does seem to be something going on where the models do seem to need to have a picture of the world and make a distinction between things that are true and things that are not true.

If you think of how the models are trained, they read a bunch of stuff on the internet. A lot of it's true. Some of it, more than we'd like, is false. And when you're training the model, it has to model all of it. And so, I think it's parsimonious, I think it's useful to the models picture of the world for it to know when things are true and for it to know when things are false.

And then the hope is, can we amplify that signal? Can we either use our internal understanding of the model as an indicator for when the model is lying, or can we use that as a hook for further training? And there are at least hooks. There are at least beginnings of how to try to address this problem.

EZRA KLEIN: So I try as best I can, as somebody not well-versed in the technology here, to follow this work on what you're describing, which I think, broadly speaking, is interpretability, right? Can we know what is happening inside the model? And over the past year, there have been some much hyped breakthroughs in interpretability.

And when I look at those breakthroughs, they are getting the vaguest possible idea of some relationships happening inside the statistical architecture of very toy models built at a fraction of a fraction of a fraction of a fraction of a fraction of the complexity of Claude 1 or GPT-1, to say nothing of Claude 2, to say nothing of Claude 3, to say nothing of Claude Opus, to say nothing of Claude 4, which will come whenever Claude 4 comes.

We have this quality of like maybe we can imagine a pathway to interpreting a model that has a cognitive complexity of an inchworm. And meanwhile, we're trying to create a superintelligence. How do you feel about that? How should I feel about that? How do you think about that?

DARIO AMODEI: I think, first, on interpretability, we are seeing substantial progress on being able to characterize, I would say, maybe the generation of models from six months ago. I think it's not hopeless, and we do see a path. That said, I share your concern that the field is progressing very quickly relative to that.

And we're trying to put as many resources into interpretability as possible. We've had one of our co-founders basically founded the field of interpretability. But also, we have to keep up with the market. So all of it's very much a dilemma, right? Even if we stopped, then there's all these other companies in the U.S.. And even if some law stopped all the companies in the U.S., there's a whole world of this.

EZRA KLEIN: Let me hold for a minute on the question of the competitive dynamics because before we leave this question of the machines that bullshit. It makes me think of this podcast we did a while ago with Demis Hassabis, who's the head of Google DeepMind, which created AlphaFold.

And what was so interesting to me about AlphaFold is they built this system, that because it was limited to protein folding predictions, it was able to be much more grounded. And it was even able to create these uncertainty predictions, right? You know, it's giving you a prediction, but it's also telling you whether or not it is — how sure it is, how confident it is in that prediction.

That's not true in the real world, right, for these super general systems trying to give you answers on all kinds of things. You can't confine it that way. So when you talk about these future breakthroughs, when you talk about this system that would be much better at sorting truth from fiction, are you talking about a system that looks like the ones we have now, just much bigger, or are you talking about a system that is designed quite differently, the way AlphaFold was?

DARIO AMODEI: I am skeptical that we need to do something totally different. So I think today, many people have the intuition that the models are sort of eating up data that's been gathered from the internet, code repos, whatever, and kind of spitting it out intelligently, but sort of spitting it out. And sometimes that leads to the view that the models can't be better than the data they're trained on or kind of can't figure out anything that's not in the data they're trained on. You're not going to get to Einstein level physics or Linus Pauling level chemistry or whatever.

I think we're still on the part of the curve where it's possible to believe that, although I think we're seeing early indications that it's false. And so, as a concrete example of this, the models that we've trained, like Claude 3 Opus, something like 99.9 percent accuracy, at least the base model, at adding 20-digit numbers. If you look at the training data on the internet, it is not that accurate at adding 20-digit numbers. You'll find inaccurate arithmetic on the internet all the time, just as you'll find inaccurate political views. You'll find inaccurate technical views. You're just going to find lots of inaccurate claims.

But the models, despite the fact that they're wrong about a bunch of things, they can often perform better than the average of the data they see by — I don't want to call it averaging out errors, but there's some underlying truth, like in the case of arithmetic. There's some underlying algorithm used to add the numbers.

And it's simpler for the models to hit on that algorithm than it is for them to do this complicated thing of like, OK, I'll get it right 90 percent of the time and wrong 10 percent of the time, right? This connects to things like Occam's razor and simplicity and parsimony in science. There's some relatively simple web of truth out there in the world, right?

We were talking about truth and falsehood and bullshit. One of the things about truth is that all the true things are connected in the world, whereas lies are kind of disconnected and don't fit into the web of everything else that's true.

[MUSIC PLAYING]

EZRA KLEIN: So if you're right and you're going to have these models that develop this internal web of truth, I get how that model can do a lot of good. I also get how that model could do a lot of harm. And it's not a model, not an A.I. system I'm optimistic that human beings are going to understand at a very deep level, particularly not when it is first developed. So how do you make rolling something like that out safe for humanity?

DARIO AMODEI: So late last year, we put out something called a responsible scaling plan. So the idea of that is to come up with these thresholds for an A.I. system being capable of certain things. We have what we call A.I. safety levels that in analogy to the biosafety levels, which are like, classify how dangerous a virus is and therefore what protocols you have to take to contain it, we're currently at what we describe as A.S.L. 2.

A.S.L. 3 is tied to certain risks around the model of misuse of biology and ability to perform certain cyber tasks in a way that could be destructive. A.S.L. 4 is going to cover things like autonomy, things like probably persuasion, which we've talked about a lot before. And at each level, we specify a certain amount of safety research that we have to do, a certain amount of tests that we have to pass. And so, this allows us to have a framework for, well, when should we slow down? Should we slow down now? What about the rest of the market?

And I think the good thing is we came out with this in September, and then three months after we came out with ours, OpenAI came out with a similar thing. They gave it a different name, but it has a lot of properties in common. The head of DeepMind at Google said, we're working on a similar framework. And I've heard informally that Microsoft might be working on a similar framework. Now, that's not all the players in the ecosystem, but you've probably thought about the history of regulation and safety in other industries maybe more than I have.

This is the way you get to a workable regulatory regime. The companies start doing something, and when a majority of them are doing something, then government actors can have the confidence to say, well, this won't kill the industry. Companies are already engaging in this. We don't have to design this from scratch. In many ways, it's already happening.

And we're starting to see that. Bills have been proposed that look a little bit like our responsible scaling plan. That said, it kind of doesn't fully solve the problem of like, let's say we get to one of these thresholds and we need to understand what's going on inside the model. And we don't, and the prescription is, OK, we need to stop developing the models for some time.

If it's like, we stop for a year in 2027, I think that's probably feasible. If it's like we need to stop for 10 years, that's going to be really hard because the models are going to be built in other countries. People are going to break the laws. The economic pressure will be immense.

So I don't feel perfectly satisfied with this approach because I think it buys us some time, but we're going to need to pair it with an incredibly strong effort to understand what's going on inside the models.

EZRA KLEIN: To the people who say, getting on this road where we are barreling towards very powerful systems is dangerous — we shouldn't do it at all, or we shouldn't do it this fast — you have said, listen, if we are going to learn how to make these models safe, we have to make the models, right? The construction of the model was meant to be in service, largely, to making the model safe.

Then everybody starts making models. These very same companies start making fundamental important breakthroughs, and then they end up in a race with each other. And obviously, countries end up in a race with other countries. And so, the dynamic that has taken hold is there's always a reason that you can justify why you have to keep going.

And that's true, I think, also at the regulatory level, right? I mean, I do think regulators have been thoughtful about this. I think there's been a lot of interest from members of Congress. I talked to them about this. But they're also very concerned about the international competition. And if they weren't, the national security people come and talk to them and say, well, we definitely cannot fall behind here.

And so, if you don't believe these models will ever become so powerful, they become dangerous, fine. But because you do believe that, how do you imagine this actually playing out?

DARIO AMODEI: Yeah, so basically, all of the things you've said are true at once, right? There doesn't need to be some easy story for why we should do X or why we should do Y, right? It can be true at the same time that to do effective safety research, you need to make the larger models, and that if we don't make models, someone less safe will. And at the same time, we can be caught in this bad dynamic at the national and international level. So I think of those as not contradictory, but just creating a difficult landscape that we have to navigate.

Look, I don't have the answer. Like, I'm one of a significant number of players trying to navigate this. Many are well-intentioned, some are not. I have a limited ability to affect it. And as often happens in history, things are often driven by these kind of impersonal pressures. But one thought I have and really want to push on with respect to the R.S.P.s —

EZRA KLEIN: Can you say what the R.S.P.s are?

DARIO AMODEI: Responsible Scaling Plan, the thing I was talking about before. The levels of A.I. safety, and in particular, tying decisions to pause scaling to the measurement of specific dangers or the absence of the ability to show safety or the presence of certain capabilities. One way I think about it is, at the end of the day, this is ultimately an exercise in getting a coalition on board with doing something that goes against economic pressures.

And so, if you say now, 'Well, I don't know. These things, they might be dangerous in the future. We're on this exponential.' It's just hard. Like, it's hard to get a multi-trillion dollar company. It's certainly hard to get a military general to say, all right, well, we just won't do this. It'll confer some huge advantage to others. But we just won't do this.

I think the thing that could be more convincing is tying the decision to hold back in a very scoped way that's done across the industry to particular dangers. My testimony in front of Congress, I warned about the potential misuse of models for biology. That isn't the case today, right? You can get a small uplift to the models relative to doing a Google search, and many people dismiss the risk. And I don't know — maybe they're right. The exponential scaling laws suggest to me that they're not right, but we don't have any direct hard evidence.

But let's say we get to 2025, and we demonstrate something truly scary. Most people do not want technology out in the world that can create bioweapons. And so I think, at moments like that, there could be a critical coalition tied to risks that we can really make concrete. Yes, it will always be argued that adversaries will have these capabilities as well. But at least the trade-off will be clear, and there's some chance for sensible policy.

I mean to be clear, I'm someone who thinks the benefits of this technology are going to outweigh its costs. And I think the whole idea behind RSP is to prepare to make that case, if the dangers are real. If they're not real, then we can just proceed and make things that are great and wonderful for the world. And so, it has the flexibility to work both ways.

Again, I don't think it's perfect. I'm someone who thinks whatever we do, even with all the regulatory framework, I doubt we can slow down that much. But when I think about what's the best way to steer a sensible course here, that's the closest I can think of right now. Probably there's a better plan out there somewhere, but that's the best thing I've thought of so far.

EZRA KLEIN: One of the things that has been on my mind around regulation is whether or not the founding insight of Anthropic of OpenAI is even more relevant to the government, that if you are the body that is supposed to, in the end, regulate and manage the safety of societal-level technologies like artificial intelligence, do you not need to be building your own foundation models and having huge collections of research scientists and people of that nature working on them, testing them, prodding them, remaking them, in order to understand the damn thing well enough — to the extent any of us or anyone understands the damn thing well enough — to regulate it?

I say that recognizing that it would be very, very hard for the government to get good enough that it can build these foundation models to hire those people, but it's not impossible. I think right now, it wants to take the approach to regulating A.I. that it somewhat wishes it took to regulating social media, which is to think about the harms and pass laws about those harms earlier.

But does it need to be building the models itself, developing that kind of internal expertise, so it can actually be a participant in different ways, both for regulatory reasons and maybe for other reasons, for public interest reasons? Maybe it wants to do things with a model that they're just not possible if they're dependent on access to the OpenAI, the Anthropic, the Google products.

DARIO AMODEI: I think government directly building the models, I think that will happen in some places. It's kind of challenging, right? Like, government has a huge amount of money, but let's say you wanted to provision $100 billion to train a giant foundation model. The government builds it. It has to hire people under government hiring rules. There's a lot of practical difficulties that would come with it.

Doesn't mean it won't happen or it shouldn't happen. But something that I'm more confident of that I definitely think is that government should be more involved in the use and the finetuning of these models, and that deploying them within government will help governments, especially the U.S. government, but also others, to get an understanding of the strengths and weaknesses, the benefits and the dangers. So I'm super supportive of that.

I think there's maybe a second thing you're getting at, which I've thought about a lot as a C.E.O. of one of these companies, which is, if these predictions on the exponential trend are right, and we should be humble — and I don't know if they're right or not. My only evidence is that they appear to have been correct for the last few years. And so, I'm just expecting by induction that they continue to be correct. I don't know that they will, but let's say they are. The power of these models is going to be really quite incredible.

And as a private actor in charge of one of the companies developing these models, I'm kind of uncomfortable with the amount of power that that entails. I think that it potentially exceeds the power of, say, the social media companies maybe by a lot.

You know, occasionally, in the more science fictiony world of A.I. and the people who think about A.I. risk, someone will ask me like, OK, let's say you build the A.G.I. What are you going to do with it? Will you cure the diseases? Will you create this kind of society?

And I'm like, who do you think you're talking to? Like a king? I just find that to be a really, really disturbing way of conceptualizing running an A.I. company. And I hope there are no companies whose C.E.O.s actually think about things that way.

I mean, the whole technology, not just the regulation, but the oversight of the technology, like the wielding of it, it feels a little bit wrong for it to ultimately be in the hands — maybe I think it's fine at this stage, but to ultimately be in the hands of private actors. There's something undemocratic about that much power concentration.

EZRA KLEIN: I have now, I think, heard some version of this from the head of most of, maybe all of, the A.I. companies, in one way or another. And it has a quality to me of, Lord, grant me chastity but not yet.

Which is to say that I don't know what it means to say that we're going to invent something so powerful that we don't trust ourselves to wield it. I mean, Amazon just gave you guys $2.75 billion. They don't want to see that investment nationalized.

No matter how good-hearted you think OpenAI is, Microsoft doesn't want GPT-7, all of a sudden, the government is like, whoa, whoa, whoa, whoa, whoa. We're taking this over for the public interest, or the U.N. is going to handle it in some weird world or whatever it might be. I mean, Google doesn't want that.

And this is a thing that makes me a little skeptical of the responsible scaling laws or the other iterative versions of that I've seen in other companies or seen or heard talked about by them, which is that it's imagining this moment that is going to come later, when the money around these models is even bigger than it is now, the power, the possibility, the economic uses, the social dependence, the celebrity of the founders. It's all worked out. We've maintained our pace on the exponential curve. We're 10 years in the future.

And at some point, everybody is going to look up and say, this is actually too much. It is too much power. And this has to somehow be managed in some other way. And even if the C.E.O.s of the things were willing to do that, which is a very open question by the time you get there, even if they were willing to do that, the investors, the structures, the pressure around them, in a way, I think we saw a version of this — and I don't know how much you're going to be willing to comment on it — with the sort of OpenAI board, Sam Altman thing, where I'm

very convinced that wasn't about A.I. safety. I've talked to figures on both sides of that. They all sort of agree it wasn't about A.I. safety. But there was this moment of, if you want to press the off switch, can you, if you're the weird board created to press the off switch. And the answer was no, you can't, right? They'll just reconstitute it over at Microsoft.

There's functionally no analogy I know of in public policy where the private sector built something so powerful that when it reached maximum power, it was just handed over in some way to the public interest.

DARIO AMODEI: Yeah, I mean, I think you're right to be skeptical, and similarly, what I said with the previous questions of there are just these dilemmas left and right that have no easy answer. But I think I can give a little more concreteness than what you've pointed at, and maybe more concreteness than others have said, although I don't know what others have said. We're at A.S.L. 2 in our responsible scaling plan. These kinds of issues, I think they're going to become a serious matter when we reach, say, A.S.L. 4. So that's not a date and time. We haven't even fully specified A.S.L. 4 —

EZRA KLEIN: Just because this is a lot of jargon, just, what do you specify A.S.L. 3 as? And then as you say, A.S.L. 4 is actually left quite undefined. So what are you implying A.S.L. 4 is?

DARIO AMODEI: A.S.L. 3 is triggered by risks related to misuse of biology and cyber technology. A.S.L. 4, we're working on now.

EZRA KLEIN: Be specific. What do you mean? Like, what is the thing a system could do or would do that would trigger it?

DARIO AMODEI: Yes, so for example, on biology, the way we've defined it — and we're still refining the test, but the way we've defined it is, relative to use of a Google search, there's a substantial increase in risk as would be evaluated by, say, the national security community of misuse of biology, creation of bioweapons, that either the proliferation or spread of it is greater than it was before, or the capabilities are substantially greater than it was before.

We'll probably have some more exact quantitative thing, working with folks who are ex-government biodefense folks, but something like this accounts for 20 percent of the total source of risk of biological attacks, or something increases the risk by 20 percent or something like that. So that would be a very concrete version of it. It's just, it takes us time to develop very concrete criteria. So that would be like A.S.L. 3.

A.S.L. 4 is going to be more about, on the misuse side, enabling state-level actors to greatly increase their capability, which is much harder than enabling random people. So where we would worry that North Korea or China or Russia could greatly enhance their offensive capabilities in various military areas with A.I. in a way that would give them a substantial advantage at the geopolitical level. And on the autonomy side, it's various measures of these models are pretty close to being able to replicate and survive in the wild.

So it feels maybe one step short of models that would, I think, raise truly existential questions. And so, I think what I'm saying is when we get to that latter stage, that A.S.L. 4, that is when I think it may make sense to think about what is the role of government in stewarding this technology.

Again, I don't really know what it looks like. You're right. All of these companies have investors. They have folks involved. You talk about just handing the models over. I suspect there's some way to hand over the most dangerous or societally sensitive components or capabilities of the models without fully turning off the commercial tap. I don't know that there's a solution that every single actor is happy with. But again, I get to this idea of demonstrating specific risk.

If you look at times in history, like World War I or World War II, industries' will can be bent towards the state. They can be gotten to do things that aren't necessarily profitable in the short-term because they understand that there's an emergency. Right now, we don't have an emergency. We just have a line on a graph that weirdos like me believe in and a few people like you who are interviewing me may somewhat believe in. We don't have clear and present danger.

EZRA KLEIN: When you imagine how many years away, just roughly, A.S.L. 3 is and how many years away A.S.L. 4 is, right, you've thought a lot about this exponential scaling curve. If you just had to guess, what are we talking about?

DARIO AMODEI: Yeah, I think A.S.L. 3 could easily happen this year or next year. I think A.S.L. 4 —

EZRA KLEIN: Oh, Jesus Christ.

DARIO AMODEI: No, no, I told you. I'm a believer in exponentials. I think A.S.L. 4 could happen anywhere from 2025 to 2028.

EZRA KLEIN: So that is fast.

DARIO AMODEI: Yeah, no, no, I'm truly talking about the near future here. I'm not talking about 50 years away. God grant me chastity, but not now. But "not now" doesn't mean when I'm old and gray. I think it could be near term. I don't know. I could be wrong. But I think it could be a near term thing.

EZRA KLEIN: But so then, if you think about this, I feel like what you're describing, to go back to something we talked about earlier, that there's been this step function for societal impact of A.I., the curve of the capabilities exponential, but every once in a while, something happens, ChatGPT, for instance, Midjourney with photos. And all of a sudden, a lot of people feel it. They realize what has happened and they react. They use it. They deploy it in their companies. They invest in it, whatever.

And it sounds to me like that is the structure of the political economy you're describing here. Either something happens where the bioweapon capability is demonstrated or the offensive cyber weapon capability is demonstrated, and that freaks out the government, or possibly something happens, right? Describing World War I and World War II is your examples did not actually fill me with comfort because in order to bend industry to government's will, in those cases, we had to have an actual world war. It doesn't do it that easily.

You could use coronavirus, I think, as another example where there was a significant enough global catastrophe that companies and governments and even people did things you never would have expected. But the examples we have of that happening are something terrible. All those examples end up with millions of bodies.

I'm not saying that's going to be true for A.I., but it does sound like that is a political economy. No, you can't imagine it now, in the same way that you couldn't have imagined the sort of pre and post-ChatGPT world exactly, but that something happens and the world changes. Like, it's a step function everywhere.

DARIO AMODEI: Yeah, I mean, I think my positive version of this, not to be so — to get a little bit away from the doom and gloom, is that the dangers are demonstrated in a concrete way that is really convincing, but without something actually bad happening, right? I think the worst way to learn would be for something actually bad to happen. And I'm hoping every day that doesn't happen, and we learn bloodlessly.

EZRA KLEIN: We've been talking here about conceptual limits and curves, but I do want, before we end, to reground us a little bit in the physical reality, right? I think that if you're using A.I., it can feel like this digital bits and bytes, sitting in the cloud somewhere.

But what it is in a physical way is huge numbers of chips, data centers, an enormous amount of energy, all of which does rely on complicated supply chains. And what happens if something happens between China and Taiwan, and the makers of a lot of these chips become offline or get captured? How do you think about the necessity of compute power? And when you imagine the next five years, what does that supply chain look like? How does it have to change from where it is now? And what vulnerabilities exist in it?

DARIO AMODEI: Yeah, so one, I think this may end up being the greatest geopolitical issue of our time. And man, this relates to things that are way above my pay grade, which are military decisions about whether and how to defend Taiwan. All I can do is say what I think the implications for A.I. is. I think those implications are pretty stark. I think there's a big question of like, OK, we built these powerful models.

One, is there enough supply to build them? Two is control over that supply, a way to think about safety issues or a way to think about balance of geopolitical power. And three, if those chips are used to build data centers, where are those data centers going to be? Are they going to be in the U.S.? Are they going to be in a U.S. ally? Are they going to be in the Middle East? Are they going to be in China?

All of those have enormous implications, and then the supply chain itself can be disrupted. And political and military decisions can be made on the basis of where things are. So it sounds like an incredibly sticky problem to me. I don't know that I have any great insight on this. I mean, as a U.S. citizen and someone who believes in democracy, I am someone who hopes that we can find a way to build data centers and to have the largest quantity of chips available in the U.S. and allied democratic countries.

EZRA KLEIN: Well, there is some insight you should have into it, which is that you're a customer here, right? And so, five years ago, the people making these chips did not realize what the level of demand for them was going to be. I mean, what has happened to Nvidia's stock prices is really remarkable.

But also what is implied about the future of Nvidia's stock prices is really remarkable. Rana Foroohar, the Financial Times, cited this market analysis. It would take 4,500 years for Nvidia's future dividends to equal its current price, 4,500 years. So that is a view about how much Nvidia is going to be making in the next couple of years. It is really quite astounding.

I mean, you're, in theory, already working on or thinking about how to work on the next generation of Claude. You're going to need a lot of chips for that. You're working with Amazon. Are you having trouble getting the amount of compute that you feel you need? I mean, are you already bumping up against supply constraints? Or has the supply been able to change, to adapt to you?

DARIO AMODEI: We've been able to get the compute that we need for this year, I suspect also for next year as well. I think once things get to 2026, 2027, 2028, then the amount of compute gets to levels that starts to strain the capabilities of the semiconductor industry. The semiconductor industry still mostly produces C.P.U.s, right? Just the things in your laptop, not the things in the data centers that train the

A.I. models. But as the economic value of the GPUs goes up and up and up because of the value of the A.I. models, that's going to switch over.

But you know what? At some point, you hit the limits of that or you hit the limits of how fast you can switch over. And so, again, I expect there to be a big supply crunch around data centers, around chips, and around energy and power for both regulatory and physics reasons, sometime in the next few years. And that's a risk, but it's also an opportunity. I think it's an opportunity to think about how the technology can be governed.

And it's also an opportunity, I'll repeat again, to think about how democracies can lead. I think it would be very dangerous if the leaders in this technology and the holders of the main resources were authoritarian countries. The combination of A.I. and authoritarianism, both internally and on the international stage, is very frightening to me.

EZRA KLEIN: How about the question of energy? I mean, this requires just a tremendous amount of energy. And I mean, I've seen different numbers like this floating around. It very much could be in the coming years like adding a Bangladesh to the world's energy usage. Or pick your country, right? I don't know what exactly you all are going to be using by 2028.

Microsoft, on its own, is opening a new data center globally every three days. You have — and this is coming from a Financial Times article — federal projections for 20 new gas-fired power plants in the U.S. by 2024 to 2025. There's a lot of talk about this being now a new golden era for natural gas because we have a bunch of it. There is this huge need for new power to manage all this data, to manage all this compute.

So, one, I feel like there's a literal question of how do you get the energy you need and at what price, but also a more kind of moral, conceptual question of, we have real problems with global warming. We have real problems with how much energy we're using. And here, we're taking off on this really steep curve of how much of it we seem to be needing to devote to the new A.I. race.

DARIO AMODEI: It really comes down to, what are the uses that the model is being put to, right? So I think the worrying case would be something like crypto, right? I'm someone who's not a believer that whatever the energy was that was used to mine the next Bitcoin, I think that was purely additive. I think that wasn't there before. And I'm unable to think of any useful thing that's created by that.

But I don't think that's the case with A.I. Maybe A.I. makes solar energy more efficient or maybe it solves controlled nuclear fusion, or maybe it makes geoengineering more stable or possible. But I don't think we need to rely on the long run. There are some applications where the model is doing something that used to be automated, that used to be done by computer systems. And the model is able to do it faster with less computing time, right? Those are pure wins. And there are some of those.

There are others where it's using the same amount of computing resources or maybe more computing resources, but to do something more valuable that saves labor elsewhere. Then there are cases where something used to be done by humans or in the physical world, and now it's being done by the models. Maybe it does something that previously I needed to go into the office to do that thing. And now I no longer need to go into the office to do that thing.

So I don't have to get in my car. I don't have to use the gas that was used for that. The energy accounting for that is kind of hard. You compare it to the food that the humans eat and what the energy cost of producing that.

So in all honesty, I don't think we have good answers about what fraction of the usage points one way and one fraction of the usage points to others. In many ways, how different is this from the general dilemma of, as the economy grows, it uses more energy?

So I guess, what I'm saying is, it kind of all matters how you use the technology. I mean, my kind of boring short-term answer is, we get carbon offsets for all of this stuff. But let's look beyond that to the macro question here.

EZRA KLEIN: But to take the other side of it, I mean, I think the difference, when you say this is always a question we have when we're growing G.D.P., is it's not quite. It's cliché because it's true to say that the major global warming challenge right now is countries like China and India getting richer. And we want them to get richer. It is a huge human imperative, right, a moral imperative for poor people in the world to become less poor. And if that means they use more energy, then we just need to figure out how to make that work. And we don't know of a way for that to happen without them using more energy.

Adding A.I. is not that it raises a whole different set of questions, but we're already straining at the boundaries, or maybe far beyond them, of safely what we can do energetically. Now we add in this, and so maybe some of the energy efficiency gains you're going to get in rich countries get wiped out. For this sort of uncertain payoff in the future of maybe through A.I., we figure out ways to stabilize nuclear fusion or something, right, you could imagine ways that could help, but those ways are theoretical.

And in the near term, the harm in terms of energy usage is real. And also, by the way, the harm in terms of just energy prices. It's also just tricky because all these companies, Microsoft, Amazon, I mean, they all have a lot of renewable energy targets. Now if that is colliding with their market incentives, it feels like they're running really fast towards the market incentives without an answer for how all that nets out.

DARIO AMODEI: Yeah, I mean, I think the concerns are real. Let me push back a little bit, which is, again, I don't think the benefits are purely in the future. It kind of goes back to what I said before. Like, there may be use cases now that are net energy saving, or that to the extent that they're not net energy saving, do so through the general mechanism of, oh, there was more demand for this thing.

I don't think anyone has done a good enough job measuring, in part because the applications of A.I. are so new, which of those things dominate or what's going to happen to the economy. But I don't think we should assume that the harms are entirely in the present and the benefits are entirely in the future. I think that's my only point here.

EZRA KLEIN: I guess you could imagine a world where we were, somehow or another, incentivizing uses of A.I. that were yoked to some kind of social purpose. We were putting a lot more into drug discovery, or we cared a lot about things that made remote work easier, or pick your set of public goods.

But what actually seems to me to be happening is we're building more and more and more powerful models and just throwing them out there within a terms of service structure to say, use them as long as you're not trying to politically manipulate people or create a bioweapon. Just try to figure this out, right? Try to create new stories and ask it about your personal life, and make a video game with it. And Sora comes out sooner or later. Make new videos with it. And all that is going to be very energy intensive.

I am not saying that I have a plan for yoking A.I. to social good, and in some ways, you can imagine that going very, very wrong. But it does mean that for a long time, it's like you could imagine the world you're talking about, but that would require some kind of planning that nobody is engaged in, and I don't think anybody even wants to be engaged in.

DARIO AMODEI: Not everyone has the same conception of social good. One person may think social good is this ideology. Another person — we've seen that with some of the Gemini stuff.

EZRA KLEIN: Right.

DARIO AMODEI: But companies can try to make beneficial applications themselves, right? Like, this is why we're working with cancer institutes. We're hoping to partner with ministries of education in Africa, to see if we can use the models in kind of a positive way for education, rather than the way they may be used by default. So I think individual companies, individual people, can take actions to steer or bend this towards the public good.

That said, it's never going to be the case that 100 percent of what we do is that. And so I think it's a good question. What are the societal incentives, without dictating ideology or defining the public good from on high, what are incentives that could help with this?

I don't feel like I have a systemic answer either. I can only think in terms of what Anthropic tries to do.

EZRA KLEIN: But there's also the question of training data and the intellectual property that is going into things like Claude, like GPT, like Gemini. There are a number of copyright lawsuits. You're facing some. OpenAI is facing some. I suspect everybody is either facing them now or will face them.

And a broad feeling that these systems are being trained on the combined intellectual output of a lot of different people — the way that Claude can quite effectively mimic the way I write is it has been trained, to some degree, on my writing, right? So it actually does get my stylistic tics quite well. You seem great, but you haven't sent me a check on that. And this seems like somewhere where there is real liability risk for the industry. Like, what if you do actually have to compensate the people who this is being trained on? And should you?

And I recognize you probably can't comment on lawsuits themselves, but I'm sure you've had to think a lot about this. And so, I'm curious both how you understand it as a risk, but also how you understand it morally. I mean, when you talk about the people who invent these systems gaining a lot of power, and alongside that, a lot of wealth, well, what about all the people whose work went into them such that they can create images in a million different styles?

And I mean, somebody came up with those styles. What is the responsibility back to the intellectual commons? And not just to the commons, but to the actual wages and economic prospects of the people who made all this possible?

DARIO AMODEI: I think everyone agrees the models shouldn't be verbatim outputting copyrighted content. For things that are available on the web, for publicly available, our position — and I think there's a strong case for it — is that the training process, again, we don't think it's just hoovering up content and spitting it out, or it shouldn't be spitting it out. It's really much more like the process of how a human

learns from experiences. And so, our position that that is sufficiently transformative, and I think the law will back this up, that this is fair use.

But those are narrow legal ways to think about the problem. I think we have a broader issue, which is that regardless of how it was trained, it would still be the case that we're building more and more general cognitive systems, and that those systems will create disruption. Maybe not necessarily by one for one replacing humans, but they're really going to change how the economy works and which skills are valued. And we need a solution to that broad macroeconomic problem, right?

As much as I've asserted the narrow legal points that I asserted before, we have a broader problem here, and we shouldn't be blind to that. There's a number of solutions. I mean, I think the simplest one, which I recognize doesn't address some of the deeper issues here, is things around the kind of guaranteed basic income side of things.

But I think there's a deeper question here, which is like as A.I. systems become capable of larger and larger slices of cognitive labor, how does society organize itself economically? How do people find work and meaning and all of that?

And just as kind of we transition from an agrarian society to an industrial society and the meaning of work changed, and it was no longer true that 99 percent of people were peasants working on farms and had to find new methods of economic organization, I suspect there's some different method of economic organization that's going to be forced as the only possible response to disruptions to the economy that will be small at first, but will grow over time, and that we haven't worked out what that is. We need to find something that allows people to find meaning that's humane and that maximizes our creativity and potential and flourishing from A.I.

And as with many of these questions, I don't have the answer to that. Right? I don't have a prescription. But that's what we somehow need to do.

EZRA KLEIN: But I want to sit in between the narrow legal response and the broad "we have to completely reorganize society" response, although I think that response is actually possible over the decades. And in the middle of that is a more specific question. I mean, you could even take it from the instrumental side. There is a lot of effort now to build search products that use these systems, right? ChatGPT will use Bing to search for you.

And that means that the person is not going to Bing and clicking on the website where ChatGPT is getting its information and giving that website an advertising impression that they can turn into a very small amount of money, or they're not going to that website and having a really good experience with that website and becoming maybe likelier to subscribe to whoever is behind that website.

And so, on the one hand, that seems like some kind of injustice done to the people creating the information that these systems are using. I mean, this is true for perplexity. It's true for a lot of things I'm beginning to see around where the A.I.s are either trained on or are using a lot of data that people have generated at some real cost. But not only are they not paying people for that, but they're actually stepping into the middle of where they would normally be a direct relationship and making it so that relationship never happens.

That also, I think, in the long run, creates a training data problem, even if you just want to look at it instrumentally, where if it becomes nonviable to do journalism or to do a lot of things to create high quality information out there, the A.I.'s ability, right, the ability of all of your companies to get high quality, up-to-date, constantly updated information becomes a lot trickier. So there both seems to me to be both a moral and a self-interested dimension to this.

DARIO AMODEI: Yeah, so I think there may be business models that work for everyone, not because it's illegitimate to train on open data from the web in a legal sense, but just because there may be business models here that kind of deliver a better product. So things I'm thinking of are like newspapers have archives. Some of them aren't publicly available. But even if they are, it may be a better product, maybe a better experience, to, say, talk to this newspaper or talk to that newspaper.

It may be a better experience to give the ability to interact with content and point to places in the content, and every time you call that content, to have some kind of business relationship with the creators of that content. So there may be business models here that propagate the value in the right way, right? You talk about LLMs using search products. I mean, sure, you're going around the ads, but there's no reason it can't work in a different way, right?

There's no reason that the users can't pay the search A.P.I.s, instead of it being paid through advertising, and then have that propagate through to wherever the original mechanism is that paid the creators of the content. So when value is being created, money can flow through.

EZRA KLEIN: Let me try to end by asking a bit about how to live on the slope of the curve you believe we are on. Do you have kids?

DARIO AMODEI: I'm married. I do not have kids.

EZRA KLEIN: So I have two kids. I have a two-year-old and a five-year-old. And particularly when I'm doing A.I. reporting, I really do sit in bed at night and think, what should I be doing here with them? What world am I trying to prepare them for? And what is needed in that world that is different from what is needed in this world, even if I believe there's some chance — and I do believe there's some chance — that all the things you're saying are true. That implies a very, very, very different life for them.

I know people in your company with kids. I know they are thinking about this. How do you think about that? I mean, what do you think should be different in the life of a two-year-old who is living through the pace of change that you are telling me is true here? If you had a kid, how would this change the way you thought about it?

DARIO AMODEI: The very short answer is, I don't know, and I have no idea, but we have to try anyway, right? People have to raise kids, and they have to do it as best they can. An obvious recommendation is just familiarity with the technology and how it works, right? The basic paradigm of, I'm talking to systems, and systems are taking action on my behalf, obviously, as much familiarity with that as possible is, I think, helpful.

In terms of what should children learn in school, what are the careers of tomorrow, I just truly don't know, right? You could take this to say, well, it's important to learn STEM and programming and A.I. and all of that. But A.I. will impact that as well, right? I don't think any of it is going to —

EZRA KLEIN: Possibly first.

DARIO AMODEI: Yeah, right, possibly first.

EZRA KLEIN: It seems better at coding than it is at other things.

DARIO AMODEI: I don't think it's going to work out for any of these systems to just do one for one what humans are going to do. I don't really think that way. But I think it may fundamentally change industries and professions one by one in ways that are hard to predict. And so, I feel like I only have clichés here. Like get familiar with the technology. Teach your children to be adaptable, to be ready for a world that changes very quickly. I wish I had better answers, but I think that's the best I got.

EZRA KLEIN: I agree that's not a good answer. [LAUGHS] Let me ask that same question a bit from another direction, because one thing you just said is get familiar with the technology. And the more time I spend with the technology, the more I fear that happening. What I see when people use A.I. around me is that the obvious thing that technology does for you is automate the early parts of the creative process.

The part where you're supposed to be reading something difficult yourself? Well, the A.I. can summarize it for you. The part where you're supposed to sit there with a blank page and write something? Well, the A.I. can give you a first draft. And later on, you have to check it and make sure it actually did what you wanted it to do and fact-checking it. And but I believe a lot of what makes humans good at thinking comes in those parts.

And I am older and have self-discipline, and maybe this is just me hanging on to an old way of doing this, right? You could say, why use a calculator from this perspective. But my actual worry is that I'm not sure if the thing they should do is use A.I. a lot or use it a little.

This, to me, is actually a really big branching path, right? Do I want my kids learning how to use A.I. or being in a context where they're using it a lot, or actually, do I want to protect them from it as much as I possibly could so they develop more of the capacity to read a book quietly on their own or write a first draft? I actually don't know. I'm curious if you have a view on it.

DARIO AMODEI: I think this is part of what makes the interaction between A.I. and society complicated where it's sometimes hard to distinguish when is an A.I. doing something, saving you labor or drudge work, versus kind of doing the interesting part. I will say that over and over again, you'll get some technological thing, some technological system that does what you thought was the core of what you're doing, and yet, what you're doing turns out to have more pieces than you think it does and kind of add up to more things, right?

It's like before, I used to have to ask for directions. I got Google Maps to do that. And you could worry, am I too reliant on Google Maps? Do I forget the environment around me? Well, it turns out, in some ways, I still need to have a sense of the city and the environment around me. It just kind of reallocates the space in my brain to some other aspect of the task.

And I just kind of suspect — I don't know. Internally, within Anthropic, one of the things I do that helps me run the company is, I'll write these documents on strategy or just some thinking in some direction that others haven't thought. And of course, I sometimes use the internal models for that. And I think what I found is like, yes, sometimes they're a little bit good at conceptualizing the idea, but the actual genesis of the idea, I've just kind of found a workflow where I don't use them for that. They're not that helpful for that. But they're helpful in figuring out how to phrase a certain thing or how to refine my ideas.

So maybe I'm just saying — I don't know. You just find a workflow where the thing complements you. And if it doesn't happen naturally, it somehow still happens eventually. Again, if the systems get general enough, if they get powerful enough, we may need to think along other lines. But in the short-term, I, at least, have always found that. Maybe that's too sanguine. Maybe that's too optimistic.

EZRA KLEIN: I think, then, that's a good place to end this conversation. Though, obviously, the exponential curve continues. So always our final question — what are three books you'd recommend to the audience?

DARIO AMODEI: So, yeah, I've prepared three. They're all topical, though, in some cases, indirectly so. The first one will be obvious. It's a very long book. The physical book is very thick, but "The Making of the Atomic Bomb," Richard Rhodes. It's an example of technology being developed very quickly and with very broad implications. Just looking through all the characters and how they reacted to this and how people who were basically scientists gradually realized the incredible implications of the technology and how it would lead them into a world that was very different from the one they were used to.

My second recommendation is a science fiction series, "The Expanse" series of books. So I initially watched the show, and then I read all the books. And the world it creates is very advanced. In some cases, it has longer life spans, and humans have expanded into space. But we still face some of the same geopolitical questions and some of the same inequalities and exploitations that exist in our world, are still present, in some cases, worse.

That's all the backdrop of it. And the core of it is about some fundamentally new technological object that is being brought into that world and how everyone reacts to it, how governments react to it, how individual people react to it, and how political ideologies react to it. And so, I don't know. When I read that a few years ago, I saw a lot of parallels.

And then my third recommendation would be actually "The Guns of August," which is basically a history of how World War I started. The basic idea that crises happen very fast, almost no one knows what's going on. There are lots of miscalculations because there are humans at the center of it, and kind of, we somehow have to learn to step back and make wiser decisions in these key moments. It's said that Kennedy read the book before the Cuban Missile Crisis. And so I hope our current policymakers are at least thinking along the same terms because I think it is possible similar crises may be coming our way.

EZRA KLEIN: Dario Amodei, thank you very much.

DARIO AMODEI: Thank you for having me.

[MUSIC PLAYING]

EZRA KLEIN: This episode of "The Ezra Klein Show" was produced by Rollin Hu. Fact-checking by Michelle Harris. Our senior engineer is Jeff Geld. Our senior editor is Claire Gordon. The show's production team also includes Annie Galvin, Kristin Lin and Aman Sahota. Original music by Isaac Jones. Audience strategy by Kristina Samulewski and Shannon Busta. The executive producer of New York Times Opinion Audio is Annie-Rose Strasser. Special thanks to Sonia Herrero.