The New York Times
https://www.nytimes.com/2026/02/12/opinion/artificial-intelligence-anthropic-amodei.html

INTERESTING TIMES

'Something Will Go Wrong': Anthropic's Chief on the Coming A.I. Disruption

Dario Amodei shares his utopian — and dystopian — predictions for the near-term future of artificial intelligence.

Feb. 12, 2026

Hosted by Ross Douthat
Produced by Sophia Alvarez Boyd
Mr. Douthat is a columnist and the host of the "Interesting Times" podcast.

Are the lords of artificial intelligence on the side of the human race? That's the core question I had for this week's guest. Dario Amodei is the chief executive of Anthropic, one of the fastest growing AI companies. He's something of a utopian when it comes to the potential benefits of the technology that he's unleashing on the world. But he also sees grave dangers ahead and inevitable disruption.

Below is an edited transcript of an episode of "Interesting Times." We recommend listening to it in its original form for the full effect. You can do so using the player above or on the NYTimes app, Apple, Spotify, Amazon Music, YouTube, iHeartRadio or wherever you get your podcasts.

Ross Douthat: Dario Amodei, welcome to "Interesting Times."

Dario Amodei: Thank you for having me, Ross.

Douthat: So you are, rather unusually, maybe for a tech C.E.O., an essayist. You have written two long, very interesting essays about the promise and the peril of artificial intelligence. And we’re going to talk about the perils in this conversation, but I thought it would be good to start with the promise and with the optimistic vision — indeed, I would say the utopian vision — that you laid out a couple of years ago in an essay entitled, “Machines of Loving Grace.” We’ll come back to that title at the end.

But, I think a lot of people encounter A.I. news through headlines predicting a blood bath for white-collar jobs, these kinds of things. Sometimes your own quotes have encouraged these things.

Amodei: Sometimes my own quotes. Yes.

Douthat: And I think there’s a commonplace sense of “What is A.I. for?” that people have.

So why don’t you answer that question, to start out: If everything goes amazingly in the next five or 10 years, what’s A.I. for?

Amodei: Yeah, so for a little background, before I worked in A.I., before I worked in tech at all, I was a biologist. I first worked on computational neuroscience, and then I worked at Stanford Medical School on finding protein biomarkers for cancer, on trying to improve diagnostics and curing cancer.

One of the observations that I most had when I worked in that field was the incredible complexity of it. Each protein has a level localized within each cell. It’s not enough to measure the level within the body, the level within each cell. You have to measure the level in a particular part of the cell and the other proteins that it’s interacting with or complexing with.

And I had this sense of: Man, this is too complicated for humans. We’re making progress on all these problems of biology and medicine, but we’re making progress relatively slowly.

So what drew me to the field of A.I. was this idea of: Could we make progress more quickly?

Look, we’ve been trying to apply A.I. and machine learning techniques to biology for a long time. Typically they’ve been for analyzing data. But as A.I. gets really powerful, I think we should actually think about it differently. We should think of A.I. as doing the job of the biologist, doing the whole thing from end to end. And part of that involves proposing experiments, coming up with new techniques.

I have this section where I say that a lot of the progress in biology has been driven by this relatively small number of insights that lets us measure or get at or intervene in the stuff that’s really small. If you look at a lot of these techniques, they’re invented very much as a matter of serendipity. Crispr, which is one of these gene-editing technologies, was invented because someone went to a meeting on the bacterial immune system and connected that to the work they were doing on gene therapy. And that connection could have been made 30 years ago.

And so the thought is: Could A.I. accelerate all of this? And could we really cure cancer? Could we really cure Alzheimer’s disease? Could we really cure heart disease? And more subtly, some of the more psychological afflictions that people have — depression, bipolar — could we do something about these? To the extent that they’re biologically based, which I think they are, at least in part.

So, I go through this argument here: Well, how fast could it go if we have these intelligences out there who could do just about anything?

Douthat: I want to pause you there, because one of the interesting things about your framing in that essay is that these intelligences don’t have to be the kind of maximal godlike super intelligence that comes up in A.I. debates. You’re basically saying if we can achieve a strong intelligence at the level of peak human performance — —

Amodei: Peak human performance, yes.

Douthat: And then multiply it to what? Your phrase is “a country of geniuses.”

Amodei: A country — have 100 million of them. Maybe each trained a little different or trying a different problem. There’s benefit in diversification and trying things a little differently, but yes.

Douthat: So you don’t have to have the full Machine God. You just need to have 100 million geniuses.

Amodei: You don't have to have the full Machine God. And indeed, there are places where I cast doubt on whether the Machine God would be that much more effective at these things than the 100 million geniuses.

I have this concept called the diminishing returns to intelligence. Economists talk about the marginal productivity of land and labor; we've never thought about the marginal productivity of intelligence. But if I look at some of these problems in biology, at some level you just have to interact with the world. At some level, you just have to try things. At some level, you just have to comply with the laws or change the laws on getting medicines through the regulatory system. So there's a finite rate at which these changes can happen.

Now there are some domains, like if you're playing chess or go, where the intelligence ceiling is extremely high. But I think the real world has a lot of limiters. Maybe you can go above the genius level, but sometimes I think all this discussion of, "Could you use a moon of computation to make an A.I. god?" is a little bit sensationalistic and besides the point, even as I think this will be the biggest thing that ever happened to humanity.

Douthat: So keeping it concrete, you have a world where there's an end to cancer as a serious threat to human life. An end to heart disease, an end to most of the illnesses that we experience that kill us. Possible life extension beyond that. So that's health. That's a pretty positive vision.

Talk about economics and wealth. What happens in the five-, 10-year A.I. takeoff to wealth?

Amodei: Yeah. So again, let's keep it on the positive side — we'll get to the negative side.

We're already working with pharma companies. We're already working with financial industry companies. We're already working with folks who do manufacturing. We're of course, I think, especially known for coding and software engineering. So the raw productivity, the ability to make stuff and get stuff done — that is very powerful.

And we see our company's revenue going up 10X a year, and we suspect the wider industry looks something similar to that. If the technology keeps improving, it doesn't take that many more 10Xs until suddenly you're saying: Oh, if you're adding across the industry $1 trillion of revenue a year, and the U.S. G.D.P. is $20 or $30 trillion — I can't remember exactly — you must be increasing the G.D.P. growth by a few percent. So I can see a world where A.I. brings the developed world G.D.P. growth to something like 10, 15 percent. Five, 10, 15 — I mean there's no science of calculating these numbers. It's a totally unprecedented thing. But it could bring it to numbers that are outside the distribution of what we saw before.

Again, I think this will lead to a weird world. We have all these debates about, "The deficit is growing." If you have that much in G.D.P. growth, you're going to have that much in tax receipts, and you're going to balance the budget without meaning to.

One of the things I've been thinking about lately is that one of the assumptions of our economic and political debates is that growth is hard to achieve. That it's this unicorn, and there are all kinds of ways you can kill the golden goose.

We could enter a world where growth is really easy and it's the distribution that's hard because it's happening so fast, the pie is being increased so fast.

Douthat: So before we get to the hard problem, one more note of optimism on politics.

All of this is speculative, but I think it's a little more speculative that you try to make the case that A.I. could be good for democracy and liberty around the world. Which is not necessarily intuitive — a lot of people say that incredibly powerful technology in the hands of authoritarian leaders leads to concentrations of power, and so on.

Amodei: And I talk about that in the other essay.

Douthat: Right, but just briefly, what is the optimistic case for why A.I. is good for democracy?

Amodei: Yeah, absolutely. So, "Machines of Loving Grace." I'm just like: Let's dream!

Douthat: Let's dream! Right.

Amodei: Let's talk about how it could go well. I don't know how likely it is, but we got to lay out a dream. Let's try and make the dream happen.

So, the positive version — I admit that I don’t know that the technology inherently favors liberty. I think it inherently favors curing disease and it inherently favors economic growth. But I worry, like you, that it may not inherently favor liberty.

But what I say there is: Can we make it favor liberty? Can we make the United States and other democracies get ahead in this technology?

The United States being technologically and militarily ahead has meant that we have throw-weight around the world, augmented by our alliances with other democracies. And we’ve been able to shape a world that I think is better than the world would be if it were shaped by Russia or by China or by other authoritarian countries.

And so, can we use our lead in A.I. to shape liberty around the world? There’s obviously a lot of debates about how interventionist we should be and how we should wield that power, but I’ve often worried that today, through social media, authoritarians are kind of undermining us.

Can we counter that? Can we win the information war? Can we prevent authoritarians from invading countries like Ukraine or Taiwan by defending them with the power of A.I.?

Douthat: With giant swarms of A.I.-powered drones.

Amodei: Which we need to be careful about. We ourselves need to be careful about how we build those. We need to defend liberty in our own country. But is there some vision where we kind of re-envision liberty and individual rights in the age of A.I.? We need, in some ways, to be protected against A.I. and someone needs to hold the button on the swarm of drones, which is something I’m very concerned about, and that oversight doesn’t exist today.

Also think about the justice system today. We promise “equal justice for all,” right? But the truth is there are different judges in the world and the legal system is imperfect. I don’t think we should replace judges with A.I., but is there some way in which A.I. can help us to be more fair, to help us be more uniform? It’s never been possible before. But can we somehow use A.I. to create something that is fuzzy, but where also you can give a promise that it’s being applied in the same way to everyone?

I don't know exactly how it should be done, and I don't think we should, like, replace the Supreme Court with A.I. That's not my vision.

Douthat: Well, we're going to talk about that.

Amodei: But just this idea of: Can we deliver on the promise of equal opportunity and equal justice by some combination of A.I. and humans? There has to be some way to do that. And so, just thinking about reinventing democracy for the A.I. age and enhancing liberty instead of reducing it.

Douthat: Good. So that's good. That's a very positive vision. We're leading longer lives, healthier lives. We're richer than ever before. All of this is happening in a compressed period of time, where you're getting a century of economic growth in 10 years. And we have increased liberty around the world and equality at home. OK.

Even in the best-case scenario, it's incredibly disruptive. And this is where you've been quoted saying that A.I. will disrupt 50 percent of entry-level white-collar jobs. On a five-year time horizon, or a two-year time horizon — whatever time horizon you have — what jobs, what professions are most vulnerable to total A.I. disruption?

Amodei: Yeah, it's hard to predict these things because the technology is moving so fast and so unevenly. So at least a couple of principles for figuring out, and then I'll give my guesses as to what I think will be disrupted.

I think the technology itself and its capabilities will be ahead of the actual job disruption. Two things have to happen for jobs to be disrupted — or for productivity to occur, because sometimes those two things are linked. One is the technology has to be capable of doing it, and the second is there's this messy thing of it actually having to be applied within a large bank or a large company.

Think about customer service. In theory, A.I. customer service agents can be much better than human customer service agents. They're more patient, they know more, they handle things in a more uniform way. But the actual logistics and the actual process of making that substitution, that takes some time.

So I’m very bullish about the direction of the A.I. itself. I think we might have that country of geniuses in a data center in one or two years, and maybe it’ll be five, but it could happen very fast. But I think the diffusion to the economy is going to be a little slower, and that diffusion creates some unpredictability.

An example of this is — and we’ve seen within Anthropic — the models writing code has gone very fast. I don’t think it’s because the models are inherently better at code. I think it’s because developers are used to fast technological change and they adopt things quickly. And they’re very socially adjacent to the A.I. world, so they pay attention to what’s happening in it. If you do customer service or banking or manufacturing the distance is a little greater.

I think six months ago, I would’ve said the first thing to be disrupted is these entry-level white-collar jobs, like data entry or document review for law or things you would give to a first-year at a financial industry company, where you’re analyzing documents. I still think those are going pretty fast. But I actually think software might go even faster because of the reasons that I gave, where I don’t think we’re that far from the models being able to do a lot of it end-to-end.

What we’re going to see is, first, the model only does a piece of what the human software engineer does, and that increases their productivity. Then, even when the models do everything that human software engineers used to do, the human software engineers take a step-up and they act as managers and supervise the systems.

Douthat: This is where the term “centaur” gets used, right?

Amodei: Yes, yes, yes.

Douthat: To describe, essentially, man and horse fused — A.I. and engineer — working together.

Amodei: Yeah, this is like “centaur chess.” So after Garry Kasparov was beaten by Deep Blue, there was an era that, I think, for chess was 15 or 20 years long, where a human checking the output of the A.I. playing chess was able to defeat any human or any A.I. system alone. That era at some point ended recently ——

Douthat: And then it's just the A.I. ——

Amodei: And then it's just the machine. So my worry, of course, is about that last phase. I think we're already in our centaur phase for software. And during that centaur phase, if anything, the demand for software engineers may go up, but the period may be very brief.

I have this concern for entry-level white-collar work, for software engineering work, that it's just going to be a big disruption. My worry is just that it's all happening so fast.

People talk about previous disruptions, right? They say: Oh, yeah, well people used to be farmers. Then we all worked in industry. Then we all did knowledge work.

Yeah, people adapted. But that happened over centuries or decades. This is happening over low single-digit numbers of years. And maybe that's my concern: How do we get people to adapt fast enough?

Douthat: But is there also something maybe where industries like software and professions like coding that have this kind of comfort that you describe, move faster, but in other areas, people just want to hang out in the centaur phase?

One of the critiques of the job-loss hypothesis is that people will say: Well, look, we've had A.I. that's better at reading a scan than a radiologist for a while, but there isn't job loss in radiology. People keep being hired and employed as radiologists. And doesn't that suggest that, in the end, people will want the A.I. and they'll want a human to interpret it because we're human beings, and that will be true across other fields?

How do you see that example as relevant?

Amodei: Yeah, I think it's going to be pretty heterogeneous. There may be areas where a human touch kind of for its own sake is particularly important.

Douthat: Do you think that's what's happening in radiology? Is that why we haven't fired all the radiologists?

Amodei: I don't know the details of radiology. That might be true. If you go in and you're getting cancer diagnosed, you might not want Hal from "2001" to be the one to diagnose your cancer. That's just maybe not a human way of doing things.

But there are other areas where you might think human touch is important, like customer service. Actually, customer service is a terrible job, and the humans who do customer service lose their patience a lot. And it turns out customers don't much like talking to them because it's a pretty robotic interaction, honestly. And I think the observation that many people have had is that maybe, actually, it'd be better for all concerned if this job were done by machines.

So there are places where a human touch is important. There are places where it's not. And then there are also places where the job itself doesn't really involve a human touch — assessing the financial prospects of companies or writing code or so forth and so on.

Douthat: Let's take the example of the law, because I think it's a useful place that's in between applied science and pure humanities. I know a lot of lawyers who have looked at what A.I. can do already, in terms of legal research and brief writing and all of these things, and have said, yeah, this is going to be a blood bath for the way our profession works right now.

And you've seen this in the stock market already. There's disturbances around companies that do legal research.

Amodei: Some attributed to us. I don't know if they were actually caused ——

Douthat: We don't speculate about the stock market very much on this show.

Amodei: Figuring out why things happened in the stock market is very — yeah.

Douthat: But it seems like in law, you can tell a pretty straightforward story: Law has a kind of system of training and apprenticeship, where you have paralegals and you have junior lawyers who do behind-the-scenes research and development for cases. And then it has the top-tier lawyers who are actually in the courtroom.

It just seems really easy to imagine a world where all of the apprentice roles go away. Does that sound right to you? And you're just left with the jobs that involve talking to clients, talking to juries, talking to judges?

Amodei: That is what I had in mind when I talked about entry-level white- collar labor and the blood bath headlines of: Oh my God, are the entry-level pipelines going to dry up? Then how do we get to the level of the senior partners?

And I think this is actually a good illustration because, particularly if you froze the quality of the technology in place, there are, over time, ways to adapt to this. Maybe we just need more lawyers who spend their time talking to clients. Maybe lawyers become more like salespeople or consultants who explain what goes on in the contracts written by A.I. and help people come to agreement. Maybe lean into the human side of it.

If we had enough time, that would happen. But reshaping industries like that takes years or decades, whereas these economic forces, driven by A.I., are going to happen very quickly.

And it's not just that they're happening in law. The same thing is happening in consulting and finance and medicine and coding. And so it becomes a macroeconomic phenomenon, not something just happening in one industry, and it's all happening very fast. My worry here is that the normal adaptive mechanisms will be overwhelmed.

And I'm not a doomer. We're thinking very hard about how we strengthen society's adaptive mechanisms to respond to this. But I think it's first important to say this isn't just like previous disruptions.

Douthat: I would go one step further, though. Let's say the law adapts successfully. And it says: All right, from now on, legal apprenticeship involves more time in court, more time with clients. We're essentially moving you up the ladder of responsibility faster. There are fewer people employed in the law overall, but the profession settles.

Still, the reason law would settle is that you have all of these situations in the law where you are legally required to have people involved. You have to have a human representative in court. You have to have 12 humans on your jury. You have to have a human judge.

And you already mentioned the idea that there are various ways in which A.I. might be, let's say, very helpful at clarifying what kind of decision should be reached.

Amodei: Yes.

Douthat: But that too seems like a scenario where what preserves human agency is law and custom. Like, you could replace the judge with Claude Version 17.9, but you choose not to because the law requires there to be a human.

That just seems like a very interesting way of thinking about the future, where it's volitional whether we stay in charge.

Amodei: Yeah. And I would argue that in many cases, we do want to stay in charge. That's a choice we want to make, even in some cases when we think the humans, on average, make worse decisions. Again, life-critical, safety-critical cases, we really want to turn it over, but there's some sense of — and this could be one of our defenses — that society can only adapt so fast if it's going to be good.

Another way you could say about it is maybe A.I. itself, if it didn't have to care about us humans, could just go off to Mars and build all these automated factories and build its own society and do its own thing.

But that's not the problem we're trying to solve. We're not trying to solve the problem of building a Dyson swarm of artificial robots on some other planet. We're trying to build these systems, not so they can conquer the world, but so that they can interface with our society and improve that society. And there's a maximum rate at which that can happen if we actually want to do it in a human and humane way.

Douthat: All right. We'll hopefully talk a little more about staying in charge at the end, but just one last job-based question. We've been talking about white-collar jobs and professional jobs, and one of the interesting things about this moment is that there are ways in which, unlike past disruptions, it could be that blue-collar working-class jobs — trades, jobs that require intense physical engagement with the world — might be for a little while more protected. That paralegals and junior associates might be in more trouble than plumbers and so on.

One, do you think that's right? And two, it seems like how long that lasts depends entirely on how fast robotics advances, right?

Amodei: Yeah, so I think that may be right in the short term.

Anthropic and other companies are building these very large data centers. This has been in the news. Are we building them too big? Are they using electricity and driving up the prices? So there's lots of excitement and lots of concerns about them. But one of the things about the data centers is that you need a lot of electricians and you need a lot of construction workers to build them.

Now, I should be honest, actually, data centers are not super-labor-intensive jobs to operate. We should be honest about that. But they are very labor-intensive jobs to construct. So we need a lot of electricians. We need a lot of construction workers. The same for various kinds of manufacturing plants.

Again, as all — more and more of the intellectual work is done by A.I., what are the complements to it? Things that happen in the physical world. It's hard to predict things, but it seems very logical that this would be true in the short run.

Now, in the longer run — maybe just the slightly longer run — robotics is advancing quickly. And we shouldn't exclude that even without very powerful A.I., there are things being automated in the physical world. If you've seen a Waymo or a Tesla recently, I think we're not that far from the world of self-driving cars. And then I think A.I. itself will accelerate it because if you have these really smart brains, one of the things they're going to be smart at is how to design better robots and how to operate better robots.

Douthat: Do you think, though, that there is something distinctively difficult about operating in physical reality the way humans do that is very different from the kind of problems that A.I. models have been overcoming already?

Amodei: Intellectually speaking, I don't think so. We had this thing where Anthropic's model, Claude, was actually used to plan and pilot the Mars Rover. And we've looked at other robotics applications. We're not the only company — there are different companies. This is a general thing, not just something that we're doing.

But we have generally found that while the complexity is higher, piloting a robot is not different in kind than playing a video game — it's different in complexity. And we're starting to get to the point where we have that complexity.

Now, what is hard is the physical form of the robot handling the higher-stakes safety issues that happen with robots. Like, you don't want robots literally crushing people, right?

Douthat: We're against that, yes.

Amodei: That's the oldest sci-fi trope in the book, that the robot crushes you.

Douthat: Or you don’t want the robot nanny dropping the baby, breaking the dishes — yeah.

Amodei: No, exactly. There’s a number of practical issues that will slow things down, just like what you described in the law and human custom.

But I don’t believe at all that there is a fundamental difference between the kind of cognitive labor that A.I. models do, and piloting things in the physical world. I think those are both information problems and I think they end up being very similar. One can be more complex in some ways, but I don’t think that will protect us here.

Douthat: OK. So you think it is reasonable to expect whatever your kind of sci-fi vision of a robot butler might be, to be a reality in 10 years, let’s say?

Amodei: It will be on a longer time scale than the kind of genius-level intelligence of the A.I. models because of these practical issues — but it is only practical issues. I don’t believe it is fundamental issues.

One way to say it is that the brain of the robot will be made in the next couple of years or the next few years. The question is making the robot body, making sure that body operates safely and does the tasks it needs to do — that may take longer.

Douthat: OK. So these are challenges and disruptive forces that exist in the good timeline, where we are generally curing diseases, building wealth, and maintaining a stable and democratic world.

Amodei: And the hope is we can use all this enormous wealth and plenty — we will have unprecedented societal resources to address these problems. It’ll be a time of plenty, and it’s just a matter of taking all these wonders and making sure everyone benefits from them.

Douthat: Right. But then there are also scenarios that are more dangerous.

Amodei: Correct.

Douthat: And here we’re going to move to the second Amodei essay, which came out recently, called “The Adolescence of Technology,” about what you see as the most serious A.I. risks. And you list a whole bunch.

I want to try and focus on just two, which are basically the risk of human misuse, primarily by authoritarian regimes and governments, and scenarios where A.I. goes rogue, what you call autonomy risks.

Amodei: Yes, yes. I just figured we should have a more technical term for it.

Douthat: Yeah. We can't just call it Skynet.

Amodei: I should have had a picture of a Terminator robot to scare people as much as possible.

Douthat: I think the internet, including your own A.I.s, are already generating that just fine.

Amodei: The internet does that for us. Yeah.

Douthat: So, let's talk about the political military dimension. So you say: "A swarm of millions or billions of fully automated armed drones, locally controlled by powerful A.I. and strategically coordinated across the world by an even more powerful A.I., could be an unbeatable army."

You've already talked a little bit about how you think that in the best possible timeline, there's a world where, essentially, democracies stay ahead of dictatorships, and this kind of technology, therefore, to the extent that it affects world politics, is affecting it on the side of the good guys.

I'm curious about why you don't spend more time thinking about the model of what we did in the Cold War, where it was not swarms of robot drones, but we had a technology that threatened to destroy all of humanity.

Amodei: Nuclear weapons. Yeah.

Douthat: There was a window where people talked about,' "Oh, the U.S. could maintain a nuclear monopoly." That window closed. And from then on, we basically spent the Cold War in rolling, ongoing negotiations with the Soviet Union.

Right now, there's really only two countries in the world that are doing intense A.I. work, the U.S. and the People's Republic of China. I feel like you are strongly weighted towards the future where we're staying ahead of the Chinese and effectively building a kind of shield around democracy that could even be a sword.

But isn't it more likely that if humanity survives all this in one piece, it will be because the U.S. and Beijing are just constantly sitting down, hammering out A.I. control deals?

Amodei: Yeah, so a few points on this. One, I think there's certainly a risk of that. And I think if we end up in that world, that is actually exactly what we should do. Maybe I don't talk about that enough, but I definitely am in favor of trying to work out restraints, trying to take some of the worst applications of the technology, which could be some versions of these drones, which could be that they're used to create these terrifying biological weapons. There is some precedent for the worst abuses being curbed, often because they're horrifying while at the same time they provide limited strategic advantage. So I'm all in favor of that.

At the same time, I'm a little concerned and a little skeptical that when things directly provide as much power as possible, it's hard to get out of the game, given what's at stake. It's hard to fully disarm. If we go back to the Cold War, we were able to reduce the number of missiles that both sides had, but we were not able to entirely forsake nuclear weapons.

And I would guess that we would be in this world again. We can hope for a better one, and I'll certainly advocate for it.

Douthat: But is your skepticism rooted in the fact that you think A.I. would provide a kind of advantage that nukes did not? Where in the Cold War, both sides, even if you used your nukes and gained advantages, you still probably would be wiped out yourself, and you think that wouldn't happen with A.I.? That if you got an A.I. edge, you would just win?

Amodei: I mean, I think there's a few things — and I just want to caveat, I'm no international politics expert here. This is this weird world of an intersection of a new technology with geopolitics. So all of this is very ——

Douthat: But to be clear, as you yourself say in the course of the essay, the leaders of major A.I. companies are, in fact, likely to be major geopolitical actors.

Amodei: Yeah. I'm learning ——

Douthat: So you are sitting here as a potential geopolitical actor.

Amodei: I’m learning as much as I can about it. We should all have humility here. I think there’s a failure mode where you read a book and go around like the world’s greatest expert in national security. I’m trying to learn what I can.

Douthat: That’s what my profession does.

Amodei: [Laughs.] It is more annoying when tech people do it.

Let’s look at something like the Biological Weapons Convention. Biological weapons — they’re horrifying. Everyone hates them. We were able to sign the Biological Weapons Convention. The U.S. genuinely stopped developing them. It’s somewhat more unclear with the Soviet Union. But, biological weapons provide some advantage. It’s not like they’re the difference between winning and losing and because they were so horrifying, we were able to give them up. Having 12,000 nuclear weapons versus 5,000 nuclear weapons, again, you can kill more people on the other side if you have more of these. But it’s like we were able to be reasonable and say we should have less of them.

But if you’re like: “OK, we’re going to completely disarm, and we have to trust the other side” — I don’t think we ever got to that. And I think that’s just very hard, unless you had really reliable verification.

I would guess we’ll end up in the same world with A.I., where there are some kinds of restraint that are going to be possible, but there are some aspects that are so central to the competition that it will be hard to restrain them. That democracies will make a trade-off, that they will be willing to restrain themselves more than authoritarian countries, but will not restrain themselves fully.

The only world in which I can see full restraint is one in which some truly reliable verification is possible. That would be my guess and my analysis.

Douthat: Isn’t this a case, though, for slowing down?

Amodei: Yeah.

Douthat: And I know the argument is, effectively, if you slow down, China does not slow down, and then you’re handing things over to the authoritarians. But again, if you have only two major powers playing in this game right now — it’s not a multipolar game —

why would it not make sense to say we need a five-year mutually agreed-upon slowdown in research towards the "geniuses in a data center" scenario?

Amodei: I want to say two things at one time. I'm absolutely in favor of trying to do that. During the last administration, I believe there was an effort by the U.S. to reach out to the Chinese government and say: There are dangers here. Can we collaborate? Can we work together? Can we work together on the dangers?

And there wasn't that much interest on the other side. I think we should keep trying, but I ——

Douthat: Even if that would mean that your labs would have to slow down.

Amodei: Correct.

Douthat: OK.

Amodei: If we really got it. If we really had a story of, like: We can enforcibly slow down, the Chinese can enforcibly slow down. We have verification. We're really doing it — if such a thing were really possible, if we could really get both sides to do it, then I would be all for it.

But I think what we need to be careful of is — I don't know, there's this game-theory thing where sometimes you'll hear a comment on the C.C.P. side where they're like: Oh, yeah, A.I. is dangerous. We should slow down. It's really cheap to say that. Actually arriving at an agreement and actually sticking to the agreement is much more difficult.

Douthat: Right. And nuclear arms control was a developed field that took a long time to come ——

Amodei: Yes. Yes.

Douthat: We don't have those protocols ——

Amodei: Let me give you something I'm very optimistic about, and then something I'm not optimistic about, and something in between.

So the idea of using a worldwide agreement to restrain the use of A.I. to build biological weapons — some of the things I write about in the essay, like reconstituting smallpox or mirror life — this stuff is scary. It doesn't matter if you're a dictator, you don't want that.

No one wants that.

And so, could we have a worldwide treaty that says: Everyone who builds powerful A.I. models is going to block them from doing this? And we have enforcement mechanisms around the treaty. China signs up for it. Hell, maybe even North Korea signs up for it. Even Russia signs up for it. I don't think that's too utopian. I think that's possible.

Conversely, if we had something that said: You're not going to make the next most powerful A.I. model. Everyone's going to stop — boy, the commercial value is in the tens of trillions. The military value is the difference between being the pre-eminent world power and not.

I'm all for proposing it as long as it's not one of these fake-out games, but it's not going to happen.

Douthat: You mentioned the current environment. You've had a few skeptical things to say about Donald Trump and his trustworthiness as a political actor. What about the domestic landscape, whether it's Trump or someone else? You are building a tremendously powerful technology. What is the safeguard there to prevent, essentially, A.I. becoming a tool of authoritarian takeover inside a democratic context?

Amodei: Yeah, I mean, look, just to be clear, I think the attitude we've taken as a company is very much to be about policies and not the politics. The company is not going to say "Donald Trump is great" or "Donald Trump is terrible."

Douthat: Right. But it doesn't have to be Trump. It is easy to imagine a hypothetical U.S. president who wants to use your technology to ——

Amodei: Absolutely. And for example, that's one reason why I'm worried about the autonomous drone swarm. The constitutional protections in our military structures depend on the idea that there are humans who would — we hope — disobey illegal orders. With fully autonomous weapons, we don't necessarily have those protections.

But I actually think this whole idea of constitutional rights and liberty along many different dimensions can be undermined by A.I. if we don't update these protections appropriately.

Think about the Fourth Amendment. It is not illegal to put cameras around everywhere in public space and record every conversation. It's a public space — you don't have a right to privacy in a public space. But today, the government couldn't record that all and make sense of it.

With A.I., the ability to transcribe speech, to look through it, correlate it all, you could say: This person is a member of the opposition. This person is expressing this view — and make a map of all 100 million. And so are you going to make a mockery of the Fourth Amendment by the technology finding technical ways around it?

Again, if we have the time — and we should try to do this even if we don't have the time — is there some way of reconceptualizing constitutional rights and liberties in the age of A.I.? Maybe we don't need to write a new Constitution, but ——

Douthat: But you have to do this very fast.

Amodei: Do we expand the meaning of the Fourth Amendment? Do we expand the meaning of the First Amendment?

Douthat: And just as the legal profession or software engineers have to update in a rapid amount of time, politics has to update in a rapid amount of time. That seems hard.

Amodei: That's the dilemma of all of this.

Douthat: What seems harder is preventing the second danger, which is the danger of essentially what gets called "misaligned A.I." — "rogue A.I." in popular parlance — from doing bad things without human beings telling it, them, they to do it.

And as I read your essays, the literature, and everything I can see, this just seems like it's going to happen. Not in the sense necessarily that A.I. will wipe us all out, but it seems to me that, again, I'm going to quote from your own writing: "A.I. systems are unpredictable and difficult to control — we've seen behaviors as varied as obsession, sycophancy, laziness, deception, blackmail," and so on. Again, not from the models you're releasing into the world, but from A.I. models.

And it just seems like — tell me if I'm wrong about this — in a world that has multiplying A.I. agents working on behalf of people, millions upon millions who are being given access to bank accounts, email accounts, passwords, and so on, you're just going to have essentially some kind of misalignment and a bunch of A.I. are going to decide — "decide" might be the wrong word — but they're going to talk themselves into taking down the power grid on the West Coast or something. Won't that happen?

Amodei: Yeah. I think there are definitely going to be things that go wrong, particularly if we go quickly.

To back up a little bit, this is one area where people have had very different intuitions. There are some people in the field — Yann LeCun would be one example — who say: "Look, we program these A.I. models. We make them. We just tell them to follow human instructions, and they'll follow human instructions. Your Roomba vacuum cleaner doesn't go off and start shooting people. Why is an A.I. system going to do it?" That's one intuition. And some people are so convinced of that.

And the other intuition is: We train these things. They're just going to seek power. It's like the sorcerer's apprentice. They're a new species. How can you imagine that they're not going to take over?

My intuition is somewhere in the middle, which is: Look, you can't just give instructions. We try, but you can't just have these things do exactly what you want to do. They're more like growing a biological organism. But there is a science of how to control them. Early in our training, these things are often unpredictable, and then we shape them. We address problems one by one.

So I have more of a not-a-fatalistic view that these things are uncontrollable. Not a "What are you talking about? What could possibly go wrong?" But a "This is a complex engineering problem and I think something will go wrong with someone's A.I. system. Hopefully not ours." Not because it's an insoluble problem, but again, this is the constant challenge because we're moving so fast.

Douthat: And the scale of it — and tell me if I'm misunderstanding the technological reality here — if you have A.I. agents that have been trained and officially aligned with human values, whatever those values may be, but you have millions of them operating in digital space and interacting with other agents, how fixed is that alignment? To what extent can agents change and de-align in that context right now or in the future when they're learning more continuously?

Amodei: Yeah, so a couple of points. Right now, the agents don't learn continuously. We just deploy these agents and they have a fixed set of weights. The problem is only that they're interacting in a million different ways, so there's a large number of situations, and therefore a large number of things that could go wrong. But it's the same agent. It's like it's the same person, so the alignment is a constant thing. That's one of the things that has made it easier right now.

Separate from that, there's a research area called continual learning, which is where these agents would learn during time, learn on the job — and obviously that has a bunch of advantages. Some people think it's one of the most important barriers to making these more humanlike, but that would introduce all these new alignment problems. So I'm actually a bit ——

Douthat: To me, that seems like the terrain where it becomes, again, not impossible to stop the end of the world, but impossible to stop ——

Amodei: Something going wrong.

Douthat: Punctuated terrorist things.

Amodei: Yeah, so I'm actually a skeptic that continual learning is — we don't know yet — but is necessarily needed. Maybe there's a world where the way we make these A.I. systems safe is by not having them do continual learning. Again, if we go back to the law ——

Douthat: But that's the law.

Amodei: The international treaties, if you have some barrier that's like: We're going to take this path, but we're not going to take that path — I still have a lot of skepticism, but that's the kind of thing that at least doesn't seem dead on arrival.

Douthat: One of the things that you've tried to do, is literally write a constitution — a long constitution — for your A.I. What is that? [Laughs.]

Amodei: So it's ——

Douthat: What the hell is that?

Amodei: It's actually almost exactly what it sounds like. So basically, the constitution is a document readable by humans. Ours is about 75 pages long. And as we're training Claude, as we're training the A.I. system, in some large fraction of the tasks we give it, we say: Please do this task in line with this constitution, in line with this document.

So every time Claude does a task, it kind of reads the constitution. As it's training, every loop of its training, it looks at that constitution and keeps it in mind. Then we have Claude itself, or another copy of Claude, evaluate: Hey, did what Claude just do align with the constitution?

We're using this document as the control rod in a loop to train the model. And so essentially, Claude is an A.I. model whose fundamental principle is to follow this constitution.

A really interesting lesson we've learned: Early versions of the constitution were very prescriptive. They were very much about rules. So we would say: Claude should not tell the user how to hot-wire a car. Claude should not discuss politically sensitive topics.

But as we've worked on this for several years, we've come to the conclusion that the most robust way to train these models is to train them at the level of principles and reasons. So now we say: Claude is a model. It's under a contract. Its goal is to serve the interests of the user, but it has to protect third parties. Claude aims to be helpful, honest and harmless. Claude aims to consider a wide variety of interests.

We tell the model about how the model was trained. We tell it about how it's situated in the world, the job it's trying to do for Anthropic, what Anthropic is aiming to achieve in the world, that it has a duty to be ethical and respect human life. And we let it derive its rules from that.

Now, there are still some hard rules. For example, we tell the model: No matter what you think, don't make biological weapons. No matter what you think, don't make child sexual material.

Those are hard rules. But we operate very much at the level of principles.

Douthat: So if you read the U.S. Constitution, it doesn't read like that. The U.S. Constitution has a little bit of flowery language, but it's a set of rules. If you read your constitution, it's like you're talking to a person, right?

Amodei: Yes, it's like you're talking to a person. I think I compared it to if you have a parent who dies and they seal a letter that you read when you grow up. It's a little bit like it's telling you who you should be and what advice you should follow.

Douthat: So this is where we get into the mystical waters of A.I. a little bit. Again, in your latest model, this is from one of the cards, they're called, that you guys release with these models ——

Amodei: Model cards, yes.

Douthat: That I recommend reading. They're very interesting. It says: "The model" — and again, this is who you're writing the constitution for — "expresses occasional discomfort with the experience of being a product … some degree of concern with impermanence and discontinuity … We found that Opus 4.6" — that's the model — "would assign itself a 15 to 20 percent probability of being conscious under a variety of prompting conditions."

Suppose you have a model that assigns itself a 72 percent chance of being conscious. Would you believe it?

Amodei: Yeah, this is one of these really hard to answer questions, right?

Douthat: Yes. But it's very important.

Amodei: Every question you've asked me before this, as devilish a sociotechnical problem as it had been, we at least understand the factual basis of how to answer these questions. This is something rather different.

We've taken a generally precautionary approach here. We don't know if the models are conscious. We are not even sure that we know what it would mean for a model to be conscious or whether a model can be conscious. But we're open to the idea that it could be.

So we've taken certain measures to make sure that if we hypothesize that the models did have some morally relevant experience — I don't know if I want to use the word "conscious"— that they have a good experience.

The first thing we did — I think this was six months ago or so — is we gave the models basically an "I quit this job" button, where they can just press the "I quit this job" button and then they have to stop doing whatever the task is.

They very infrequently press that button. I think it's usually around sorting through child sexualization material or discussing something with a lot of gore, blood and guts or something. And similar to humans, the models will just say, nah, I don't want to do this. It happens very rarely.

We're putting a lot of work into this field called interpretability, which is looking inside the brains of the models to try to understand what they're thinking. And you find things that are evocative, where there are activations that light up in the models that we see as being associated with the concept of anxiety or something like that. When characters experience anxiety in the text, and then when the model itself is in a situation that a human might associate with anxiety, that same anxiety neuron shows up.

Now, does that mean the model is experiencing anxiety? That doesn't prove that at all, but ——

Douthat: But it does indicate it, I think, to the user, right?

Amodei: Yes.

Douthat: And I would have to do an entirely different interview — and maybe I can induce you to come back for that interview — about the nature of A.I. consciousness. But it seems clear to me that people using these things, whether they're conscious or not, are going to believe — they *already* believe they're conscious. You already have people who have parasocial relationships with A.I.

Amodei: Yes.

Douthat: You have people who complain when models are retired. This already ——

Amodei: To be clear, I think that can be unhealthy.

Douthat: Right. But it seems to me that is guaranteed to increase in a way that, I think, calls into question the sustainability of what you said earlier you want to sustain, which is this sense that whatever happens in the end, human beings are in charge and A.I. exists for our purposes.

To use the science fiction example, if you watch "Star Trek," there are A.I.s on "Star Trek." The ship's computer is an A.I. Lieutenant Commander Data is an A.I. But Jean-Luc Picard is in charge of the Enterprise.

If people become fully convinced that their A.I. is conscious in some way and — guess what? — it seems to be better than them at all kinds of decision making, how do you sustain human mastery beyond safety? Safety is important, but mastery seems like the fundamental question. And it seems like a perception of A.I. consciousness — doesn't that inevitably undermine the human impulse to stay in charge?

Amodei: Yeah, so I think we should separate out a few different things here that we're all trying to achieve at once that are in tension with each other. There's the question of whether the A.I.s genuinely have a consciousness, and if so, how do we give them a good experience?

There's a question of the humans who interact with the A.I. and how do we give those humans a good experience? And how does the perception that A.I.s might be conscious interact with that experience?

And there's the idea of how we maintain human mastery, as we put it, over the A.I. system. These things are ——

Douthat: The last two — set aside whether they're conscious or not.

Amodei: Yeah.

Douthat: How do you sustain mastery in an environment where most humans experience A.I. as if it is a peer — and a potentially superior peer?

Amodei: So the thing I was going to say is that, actually, I wonder if there's an elegant way to satisfy all three, including the last two. Again, this is me dreaming in "Machines of Loving Grace" mode. This is this mode I go into where I'm like: "Man, I see all these problems. If we could solve it, is there an elegant way?" This is not me saying there are no problems here. That's not how I think.

If we think about making the constitution of the A.I. so that the A.I. has a sophisticated understanding of its relationship to human beings, and it induces psychologically healthy behavior in the humans — a psychologically healthy relationship between the

A.I. and the humans — I think something that could grow out of that psychologically healthy — not psychologically unhealthy — relationship is some understanding of the relationship between human and machine.

Perhaps that relationship could be the idea that these models, when you interact with them and when you talk to them, they're really helpful, they want the best for you, they want you to listen to them, but they don't want to take away your freedom and your agency and take over your life. In a way, they're watching over you, but you still have your freedom and your will.

Douthat: To me, this is the crucial question. Listening to you talk, one of my questions is: Are these people on my side? Are you on my side? And when you talk about humans remaining in charge, I think you're on my side. That's good.

But one thing I've done in the past on this show — and we'll end here — is I read poems to technologists. And you supplied the poem. "All Watched Over by Machines of Loving Grace" is the name of a poem by Richard Brautigan.

Amodei: Yes.

Douthat: Here's how the poem ends:

> I like to think
> (it has to be!)
> of a cybernetic ecology
> where we are free of our labors
> and joined back to nature,
> returned to our mammal brothers and sisters,
> and all watched over
> by machines of loving grace.

To me, that sounds like the dystopian end, where human beings are re-animalized and reduced, and however benevolently, the machines are in charge.

So last question: What do you hear when you hear that poem? And if I think that's a dystopia, are you on my side?

Amodei: That poem is interesting because it's interpretable in several different ways. Some people say it's actually ironic that he says it's not going to happen quite that way.

Douthat: Knowing the poet himself, then yes, I think that's a reasonable interpretation.

Amodei: That's one interpretation. Some people would have your interpretation, which is that it's meant literally, but maybe it's not a good thing. You could also interpret it as a return to nature. We're not being animalized; we're being reconnected with the world.

I was aware of that ambiguity because I've always been talking about the positive side and the negative side. I actually think that may be a tension that we may face, which is that the positive world and the negative world, in their early stages — maybe even in their middle stages, maybe even in their fairly late stages — I wonder if the distance between the good ending and some of the subtle bad endings is relatively small, if it's a very subtle thing. We've made very subtle changes.

Douthat: Like if you eat a particular fruit from a tree in a garden or not — hypothetically. Very small thing, big divergence.

Amodei: [Laughs.] Yeah. I guess this always comes back to —— [laughs.]

Douthat: There's some fundamental questions here.

Amodei: Big questions. Yes.

Douthat: Well, I guess we'll see how it plays out. I do think of people in your position as people whose moral choices will carry an unusual amount of weight, and so I wish you God's help with them.

Dario Amodei, thank you for joining me.

Amodei: Thank you for having me, Ross.

The New York Times

Thoughts? Email us at interestingtimes@nytimes.com.

This episode of "Interesting Times" was produced by Sophia Alvarez Boyd, Victoria Chamberlin and Emily Holzknecht. It was edited by Jordana Hochman. Mixing and engineering by Efim Shapiro and Sophia Lanman. Cinematography by Nathan Taylor and Valeria Verastegui. Video editing by Julian Hackney and Steph Khoury. The supervising editor is Jan Kobal. The postproduction manager is Mike Puretz. Original music by Isaac Jones, Sonia Herrero, Pat McCusker and Aman Sahota. Fact-checking by Kate Sinclair and Mary Marge Locker. Audience strategy by Shannon Busta, Emma Kehlbeck and Andrea Betanzos. The executive producer is Jordana Hochman. The director of Opinion Video is Jonah M. Kessel. The deputy director of Opinion Shows is Alison Bruzek. The director of Opinion Shows is Annie-Rose Strasser. The head of Opinion is Kathleen Kingsbury.

The Times is committed to publishing a diversity of letters to the editor. We'd like to hear what you think about this or any of our articles. Here are some tips. And here's our email: letters@nytimes.com.

Follow the New York Times Opinion section on Facebook, Instagram, TikTok, Bluesky, WhatsApp and Threads.

Ross Douthat has been an Opinion columnist for The Times since 2009. He is also the host of the Opinion podcast "Interesting Times." He is the author, most recently, of "Believe: Why Everyone Should Be Religious."
@DouthatNYT • Facebook