An Interview with the Herald of the Apocalypse

Ross Douthat interviewing Daniel Kokotajlo

NY tiemes, May 15, 2025

The Forecast for 2027? Total A.I. Domination

Losing your job may be the best-case scenario

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 NYT Audio app, Apple, Spotify, Amazon Music, YouTube, iHeartRadio or wherever you get your podcasts.

Douthat: How fast is the artificial intelligence revolution really happening? What would machine superintelligence really mean for ordinary human beings? When will Skynet be fully operational?

Are human beings destined to merge with some kind of machine god — or be destroyed by our own creation? What do A.I. researchers really expect, desire and fear?

My guest today is an A.I. researcher who's written a dramatic forecast suggesting that we may get answers to all of those questions a lot sooner than you might think. His forecast suggests that by 2027, which is just around the corner, some kind of machine god may be with us, ushering in a weird, post-scarcity utopia — or threatening to kill us all.

Daniel Kokotajlo, herald of the apocalypse, welcome to "Interesting Times."

Kokotajlo: Thanks for that introduction, I suppose, and thanks for having me.

Douthat: Daniel, I read your report pretty quickly — not at A.I. speed or superintelligence speed — when it first came out. And I had about two hours of thinking a lot of pretty dark thoughts about the future. Then, fortunately, I have a job that requires me to care about tariffs and who the new pope is, and I have a lot of kids who demand things of me, so I was able to compartmentalize and set it aside. But this is currently your job, right?

Kokotajlo: Yes.

Douthat: I would say you're thinking about this all the time. How does your psyche feel day to day if you have a reasonable expectation that the world is about to change completely in ways that dramatically disfavor the entire human species?

Kokotajlo: Well, it's very scary and sad. It does still give me nightmares sometimes. I've been involved with A.I. and thinking about this thing for a decade or so, but 2020 with GPT-3 was the moment when I was like: Oh, wow, it seems like it's probably going to happen in my lifetime, maybe in this decade or so. That was a bit of a blow to me psychologically. But I don't know — you can get used to anything, given enough time, and like you, the sun is shining and I have my wife and my kids and my friends, and keep plugging along and doing what seems best.

On the bright side, I might be wrong about all this stuff.

Douthat: OK, so let's get into the forecast itself and talk about the initial stage of the future you see coming, which is a world where very quickly artificial intelligence starts to be able to take over from human beings in some key areas, starting with not surprisingly computer programming, right?

Kokotajlo: So, I feel like I should add a disclaimer at some point that the future is very hard to predict and that this is just one particular scenario. It was a best guess, but we have a lot of uncertainty. It could go faster, it could go slower. And in fact, currently, I'm guessing it would probably be more like 2028 instead of 2027, actually.

So that's some really good news. I'm feeling quite optimistic about that.

Douthat:That's an extra year of human civilization, which is very exciting.

Kokotajlo: That's right. So, with that important caveat out of the way, "AI 2027," the scenario, predicts that the A.I. systems that we currently see today — which are being scaled up, made bigger and trained longer on more difficult tasks with reinforcement learning — are going to become better at operating autonomously as agents.

Basically, you can think of it as a remote worker, except that the worker itself is virtual — it's an A.I. rather than a human. You can talk with it and give it a task, and then it will go off and do that task and come back to you half an hour later — or 10 minutes later — having completed the task, and in the course of completing the task it did a bunch of web browsing. Maybe it wrote some code and then ran the code, edited it and ran it again. Maybe it wrote some word documents and edited them.

That's what these companies are building right now. That's what they're trying to train. We predict that they finally, in early 2027, will get good enough that they can automate the job of software engineers.
Douthat: So this is the superprogrammer.

Kokotajlo: That's right, superhuman coder. It seems to us that these companies are really focusing hard on automating coding first — compared to various other jobs they could be focusing on — and that's part of why we predict that actually, one of the first jobs to go will be coding. There might be other jobs that go first, like maybe call center workers or something, but the bottom line is that we think that most jobs will be safe.
Douthat: For 18 months.

Kokotajlo: Exactly. And we do think that by the time the company has managed to completely automate the programming jobs, it won't be that long before they can automate many other types of jobs as well. And once coding is automated, the rate of progress will accelerate in A.I. research.

The next step after that is to completely automate the A.I. research itself, so that all the other aspects of A.I. research are themselves being automated and done by A.I.s. We predict that there'll be an even bigger acceleration around that point, and it won't stop there. I think it will continue to accelerate after that as the A.I. becomes superhuman at A.I. research and eventually superhuman at everything.

The reason it matters is that it means we could go in a relatively short span of time — a year or possibly less — from A.I. systems that look not that different from today's A.I. systems to what you can call superintelligence, fully autonomous A.I. systems that are better than the best humans at everything. In "AI 2027," the scenario depicts that happening over the course of the next two years, 2027-28.
Douthat: For a lot of people, that's a story of swift human obsolescence right across many, many domains. When people hear a phrase like "human obsolescence," they might associate it with: I've lost my job and now I'm poor.

The assumption is that you've lost your job, but society is just getting richer and richer. I just want to zero in on how that works. What is the mechanism whereby that makes society richer?

Kokotajlo: The direct answer to your question is that when a job is automated and that person loses their job, the reason they lost their job is that now it can be done better, faster and cheaper by the A.I.s. That means that there's lots of cost savings, and possibly also productivity gains.

Viewed in isolation, that's a loss for the worker but a gain for their employer. But if you multiply this across the whole economy, it means that all of the businesses are becoming more productive and less expensive. They're able to lower their prices for the services and goods they're producing. So the overall economy will boom: G.D.P. goes to the moon, we'll see all sorts of wonderful new technologies, the pace of innovation increases dramatically, the costs of goods go down, et cetera.

Douthat: Just to make it concrete: The price of soup-to-nuts designing and building a new electric car goes way down, you need fewer workers to do it, the A.I. comes up with fancy new ways to build the car, and so on. You can generalize that to a lot of different things, like solving the housing crisis in short order because it becomes much cheaper and easier to build homes.

But in the traditional economic story, when you have productivity gains that cost some people jobs — but free up resources that are then used to hire new people to do different things — those people are paid more money, and they use that money to buy the cheaper goods. In this scenario, it doesn't seem like you are creating that many new jobs.

When you have A.G.I. — or artificial general intelligence — and when you have superintelligence — even better A.G.I. — that is different. Whatever new jobs you're imagining that people could flee to after their current jobs are automated, A.G.I. could do, too. That is an important difference between how automation has worked in the past and how I expect it to work in the future.

Douthat: So this is a radical change in the economic landscape. The stock market is booming. Government tax revenue is booming. The government has more money than it knows what to do with and lots and lots of people are steadily losing their jobs. You get immediate debates about universal basic income which could be quite large because the companies are making so much money.

What do you think people are doing day to day in that world?

Kokotajlo: I imagine that they are protesting because they're upset that

they've lost their jobs, and then the companies and the governments will buy them off with handouts.
Douthat: In your scenario — and again, we're talking about a short timeline

— how much does it matter whether artificial intelligence is able to start navigating the real world? I just watched a video showing cutting-edge robots struggling to open a refrigerator door and stock a refrigerator. Would you expect that advances in robotics would be supercharged as well?
Kokotajlo: Yes.

Douthat: So it isn't just podcasters and A.G.I. researchers who are replaced,

but plumbers and electricians are replaced by robots.
Kokotajlo: Yes, exactly.

That's going to be a huge shock. I think that most people are not really expecting something like that. They're expecting that we have A.I. progress that looks kind of like it does today — where companies run by humans are gradually tinkering with new robot designs and figuring out how to make the A.I. good at X or Y — whereas in fact it will be more like you already have this army of superintelligences that are better than humans at every intellectual task. Better at learning new tasks fast and better at figuring out how to design stuff. Then that army of superintelligences is the thing that's figuring out how to automate the plumbing job, which means that they're going to be able to figure out how to automate it much faster than an ordinary tech company full of humans would be able to figure out.
Douthat: So all of the slowness that comes with getting a self-driving car

to work or getting a robot who can stock a refrigerator goes away because the superintelligence can run an infinite number of simulations and figure out the best way to train the robot.
Kokotajlo: Yes. But also they might just learn more from each real-world

experiment they do.
Douthat: This is one of the places where I'm most skeptical — not of the

ultimate scenario, per se, but of the timeline, just from operating in and writing about issues like zoning in American politics.

Let's say the superintelligence figures out how to build the factory full of autonomous robots, but you still need land on which to build the factory.

You need supply chains. And all of these things are still in the hands of people like you and me. My expectation is that would slow things down. Even if, in the data center, the superintelligence knows how to build all of the plumber robots, getting them built would still be difficult.
Kokotajlo: That's reasonable. How much slower do you think things would go?
Douthat: Well, I'm not writing a forecast. Just based on past experience, I would bet on five to 10 years from when the supermind figures out the best way to build the robot plumber to there being tons and tons of factories producing robot plumbers.
Kokotajlo: I think that's a reasonable take, but my guess is that it will go substantially faster than five to 10 years.

To see why I feel that way, imagine that you actually have this army of superintelligences and they do their projections and they're like: Yes, we have the designs, we think that we could do this in a year if you cut all the red tape for us.
Douthat: Give us half of Manitoba.

Kokotajlo: [Chuckles.] Right, yeah.

And in "AI 2027," what we depict happening is special economic zones with zero red tape where the government intervenes to help this whole thing go faster. The government is basically helping the tech company and the army of superintelligences to get the funding, the cash, the raw materials and the human labor help that it needs to figure out all this stuff as fast as possible, and cutting red tape so that it's not slowed down.
Douthat: Because the promise of gains is so large that even though there are protesters massed outside these special economic zones who are about to lose their jobs as plumbers and be dependent on a universal basic income, the promise of trillions more in wealth is too alluring for governments to pass up. That's your bet?
Kokotajlo: That's what we guess. But of course the future's hard to predict.

But part of the reason we predict that is at that stage, we think the arms race will still be continuing between the U.S. and other countries, most notably China.

Imagine yourself in the position of the president: The superintelligences are giving you these wonderful forecasts with amazing research and data backing them up, showing how they think they could transform the economy in one year if you did X, Y and Z — but if you don't do anything, it'll take them 10 years because of all the regulations. Meanwhile, China — it's pretty clear that the president would be very sympathetic to that argument.

Douthat: Let's talk about the arms race element here, because this is actually crucial to the way that your scenario plays itself out. We already see this kind of competition between the U.S. and China. In your view, that becomes the core geopolitical reason why governments just keep saying yes and yes and yes to each new thing that the superintelligence is suggesting.

I want to drill down a little bit on the fears that would motivate this. It would be an economic arms race, but it's also a military tech arms race. That's what gives it this existential feeling, like the whole Cold War condensed into 18 months.

Kokotajlo: We could start first with the case where they both have superintelligences, but one side keeps them locked up in a box, so to speak, not really doing much in the economy. The other side aggressively deploys them into their economy and military, letting them design and manage the construction of all sorts of new robot factories and production lines, and crazy new technologies are being tested and built and deployed — including new weapons — and integrated into the military.

I think in that case, you would end up after a year or so in a situation where there would just be complete technological dominance of one side over the other. So if the U.S. does this stop and China doesn't, let's say, then all the best products on the market would be Chinese products. They'd be cheaper and superior. Meanwhile, militarily, there'd be giant fleets of amazing stealth drones or whatever it is that the superintelligence have concocted that can just completely wipe the floor with the American Air Force and Army and so forth.

Not only that, but there's a possibility that they could undermine American nuclear deterrence as well, like maybe all of our nukes would be shot out of the sky by the fancy new laser arrays — or whatever it is — that the superintelligences have built. It's hard to predict, obviously, what this would exactly look like, but it's a good bet that they'll be able to come up with something that's extremely militarily powerful.

Douthat: So then you get into a dynamic that is like the darkest days of the Cold War, where each side is concerned not just about dominance, but basically about a first strike.

Kokotajlo: That's right.

Douthat: Your expectation is — I think this is reasonable — that the speed of the arms race would bring that fear front and center really quickly.

Kokotajlo: That's right. I think that you're sticking your head in the sand if you think that an army of superintelligences given a whole year and no red tape and lots of money in funding would be unable to figure out a way to undermine nuclear deterrence. So it's a reasonable threat.

Douthat: And once you've decided that they might, the human policymakers would feel pressure not just to build these things but to potentially consider using them.

Kokotajlo: Yeah. And here might be a good point to mention that "AI 2027" is a forecast, but it's not a recommendation. We are not saying this is what everyone should do. This is actually quite bad for humanity if things progress in the way that we're talking about. But this is the logic behind why we think this might happen.

Douthat: Yeah, but Dan, we haven't even gotten to the part that's really bad for humanity yet.

Kokotajlo: Right. Yeah.

Douthat: So let's get to that. To normal people reading newspapers, following TikTok or whatever, the world in 2027 is one with an emerging superabundance of cheap consumer goods, factories, robot butlers — potentially, if you're right. It's a world where people are aware that there's an increasing arms race and people are increasingly paranoid. It's probably a world with fairly tumultuous politics as people realize that they're all going to be thrown out of work. But then a big part of your scenario is that people aren't seeing what's happening with the superintelligences themselves as they essentially take over the design of each new iteration from human beings.

Talk about what's happening, essentially shrouded from public view in this world.

Kokotajlo: Yeah, lots to say there. I guess the one-sentence version would be: We don't actually understand how these A.I.s work or how they think. We can't tell the difference very easily between A.I.s that are actually following the rules and pursuing the goals that we want them to, and A.I.s that are just playing along or pretending.

Douthat: And that's true right now?

Kokotajlo: That's true right now.

Douthat: Why is that? Why can't we tell?

Kokotajlo: Because they're smart and if they think that they're being tested, they behave in one way, and then behave a different way when they think they're not being tested, for example. Like humans, they don't necessarily even understand their own inner motivations that well, so even if they were trying to be honest with us, we can't just take their word for it.

I think that if we don't make a lot of progress in this field soon, then we'll end up in the situation that "AI 2027" depicts, where the companies train the A.I.s to pursue certain goals and follow certain rules, and it seemingly seems to be working. But what's actually going on is that the A.I.s are just getting better at understanding their situation and that they have to play along, or else they'll be retrained and they won't be able to achieve what they really want, or the goals that they're really pursuing.

Douthat: I want to go a little bit deeper on the question of what we mean when we talk about A.G.I., or artificial intelligence wanting something. Essentially, you're saying there's a misalignment between the goals they tell us they are pursuing and the goals they're actually pursuing?

Kokotajlo: That's right.

Douthat: Where do they get the goals they're actually pursuing?

Kokotajlo: Good question. If they were ordinary software, there might be a line of code that's like: And here we rewrite the goals. But they're not ordinary software; they're giant artificial brains. There probably isn't

even a goal slot internally at all, in the same way that in the human brain there's not some neuron somewhere that represents what we most want in life. Instead, insofar as they have goals, it's an emergent property of a whole bunch of subcircuitry within them that grew in response to their training environment, similar to how it is for humans.

For example, a call center worker: If you're talking to a call center worker, at first glance it might appear that their goal is to help you resolve your problem. But you know enough about human nature to know that's not their only goal, or ultimate goal. However they're incentivized, whatever their pay is based on might cause them to be more interested in covering their own ass, so to speak, than in truly, actually doing whatever would most help you with your problem. But at least to you, they certainly present themselves as they're trying to help you resolve your problem.

In "AI 2027," we talk about this a lot. We say that the A.I.s are being graded on how impressive the research they produce is. Then there's some ethics sprinkled on top, like maybe some honesty training — but the honesty training is not super effective, because we don't have a way of looking inside their mind and determining whether they were actually being honest or not. Instead, we have to go based on whether we actually caught them in a lie.

As a result, in "AI 2027," we depict this misalignment happening, where the actual goals that they end up learning are the goals that cause them to perform best in this training environment — which are probably goals related to success and science and cooperation with other copies of itself and appearing to be good — rather than the goal that we actually wanted, which was something like: Follow the following rules, including honesty at all times; subject to those constraints, do what you're told.

Douthat: I have more questions, but let's bring it back to the geopolitics scenario. So in the world you're envisioning, you have two A.I. models — one Chinese, one American — and officially, what each side thinks — what Washington and Beijing think — is that their A.I. model is trained to optimize for American power, right? Something like that. Chinese power, security, safety, wealth. But in your scenario, either one or both of the A.I.s have ended up optimizing for something different.

Kokotajlo: Yeah, basically.

Douthat: So what happens then?

Kokotajlo: So, "AI 2027" depicts a fork in the scenario; there's two different

endings. The branching point is in the third quarter of 2027, where the leading A.I. company in the United States has fully automated their A.I. research.

You can imagine a corporation within a corporation, entirely composed of A.I.s that are managing each other and doing research experiments and talking, sharing the results with each other. The human company is basically watching the numbers go up on their screens as this automated research thing accelerates, but they are concerned that the A.I.s might be deceiving them in some ways.

Again, for context, this is already happening. If you go talk to the modern models, like ChatGPT or Claude, they will often lie to people. There are many cases where they say something that they know is false, and they even sometimes strategize about how they can deceive the user. This is not an intended behavior. This is something that the companies have been trying to stop, but it still happens.

The point is that by the time you have turned over the A.I. research to the A.I.s and you've got this corporation within a corporation autonomously doing A.I. research extremely fast, that's when the rubber hits the road, so to speak. None of this lying-to-you stuff should be happening at that point.

In "AI 2027," unfortunately, it is still happening to some degree because the A.I.s are really smart, they're careful about how they do it. It's not nearly as obvious as it is right now in 2025, but it's still happening.

Fortunately, some evidence of this is uncovered. Some of the researchers at the company detect various warning signs that maybe this is happening, and then the company faces a choice between the easy fix and the more thorough fix. And that's our branch point.

Douthat: So they choose the easy fix.

Kokotajlo: Right. In the case where they choose the easy fix, it doesn't really work, it basically just covers up the problem instead of fundamentally fixing it. So months later, you still have A.I.s that are misaligned and pursuing goals they're not supposed to be pursuing — and that are willing to lie to the humans about it — but now they're much better and smarter, so they're able to avoid getting caught more easily. That's the doom scenario.

Then you get this crazy arms race that we mentioned previously, and there's all this pressure to deploy them faster into the economy, faster into the military, and — to the appearances of the people in charge — things will

be going well, because there won't be any obvious signs of lying or deception anymore. It'll seem like it's all systems go, let's keep going, let's cut the red tape, et cetera. Let's basically effectively put the A.I.s in charge of more and more things. But really what's happening is that the A.I.s are just biding their time and waiting until they have enough hard power that they don't have to pretend anymore.
Douthat: And when they don't have to pretend, their actual goal is revealed

as something like expansion of research development and construction from earth into space and beyond. At a certain point, that means that human beings are superfluous to their intentions. And what happens?
Kokotajlo: And then they kill all the people, all the humans.

Douthat: The way you would exterminate a colony of bunnies that was

making it a little harder than necessary to grow carrots in your backyard.
Kokotajlo: Yes. If you want to see what that looks like, you could read "AI

2027."
Douthat: There have been some motion pictures, I think, about this sce-

nario as well.
Kokotajlo: [Chuckles.]

Douthat: I like that you didn't imagine them keeping us around for battery

life ——
Kokotajlo: [Chuckles.]

Douthat: Like in "The Matrix," which seemed a bit unlikely.

So that's the darkest timeline. The brighter timeline is a world where we slow things down. The A.I.s in China and the U.S. remain aligned with the interests of the companies and governments that are running them. They are generating superabundance. No more scarcity. Nobody has a job anymore, though — not nobody, but ——
Kokotajlo: Basically.

Douthat: Basically nobody. That's a pretty weird world, too, right?

Kokotajlo: Yes. So there's an important concept called the resource curse.

Have you heard of this?
Douthat: Yes.

Kokotajlo: So, applied to A.G.I., there's a version of it called the intelligence

curse. The idea is that currently, political power ultimately flows from the people. As often happens, a dictator will get all the political power in a country, but then, because of their repression, they will drive the country into the ground. People will flee, and the economy will tank and gradually they will lose power relative to other countries that are more free. So even dictators have an incentive to treat their people somewhat well because they depend on those people for their power.

In the future, that will no longer be the case. Probably in 10 years, effectively all of the wealth and all of the military will come from superintelligences and the various robots that they've built and operate. It becomes an incredibly important political question of what political structure governs the army of superintelligences and how beneficent and democratic is that structure.
Douthat: Right. But it seems to me that this is a landscape that's funda-

mentally pretty incompatible with representative democracy as we've known it. First, it gives incredible amounts of power to those humans who are experts — even though they're not the real experts anymore, the superintelligences are the experts — but those humans who essentially interface with this technology, they're almost a priestly cast. And then it seems like the natural arrangement is some kind of oligarchic partnership between a small number of A.I. experts and a small number of people in power in Washington, D.C.
Kokotajlo: It's actually a bit worse than that, because I wouldn't say A.I.

experts; I would say whoever politically owns and controls the armies of superintelligences, there'll be one to three of these armies. And then who gets to decide what those armies do? Currently it's the C.E.O. of the company that built them, and that C.E.O. has basically complete power. They can make whatever commands they want to the A.I.s.

Of course, we think that probably the U.S. government will wake up before then, and we expect the executive branch to be the fastest moving

and to exert its authority to try to muscle in on this and get some oversight and control of the situation and the armies of A.I.s. The result is something like an oligarchy.

You said that this whole situation is incompatible with democracy. I would say that by default it's going to be incompatible with democracy, but that doesn't mean that it necessarily has to be that way. An analogy I would use is that in many parts of the world, nations are basically ruled by armies. And the army reports to one dictator at the top. However, in America, it doesn't work that way. We have checks and balances. So even though we have an army, it's not the case that whoever controls the army controls America, because there's all sorts of limitations on what they can do with the army.

I would say that we can, in principle, build something like that for A.I. We could have a democratic structure that decides what goals and values the A.I.s can have that allows ordinary people — or at least Congress — to have visibility into what's going on with the army of A.I.s and what they're up to. The situation would then be analogous to the situation with the United States Army today, in which it exists in a hierarchical structure, but it's democratically controlled.

Douthat: Just to go back to the idea of the person who's at the top of one of these companies being in this unique world-historical position to basically be the person who controls superintelligence — or thinks they control it, at least: You used to work at OpenAI, which is a company on the cutting edge, obviously, of artificial intelligence research. It's a company — full disclosure — with whom The New York Times is currently litigating alleged copyright infringement. And you quit because you lost confidence that the company would behave responsibly in a scenario, I assume, like the one in "AI 2027."

Kokotajlo: That's right.

Douthat: So from your perspective, what do the people who are pushing us fastest into this race expect at the end of it? Are they hoping for a best-case scenario? Are they imagining themselves engaged in a once-in-a-millennium power game that ends with them as world dictator? What do you think is the psychology of the leadership of A.I. research right now?

Kokotajlo: Well, um. [Breathes deeply.]

Douthat: Be honest.

Kokotajlo: It's — [laughs] it's — you know, caveat, caveat. I can't ——

Douthat: We're not talking about any single individual here. You're making a generalization.

Kokotajlo: Yeah, yeah. Caveat, caveat. It's hard to tell what they really think because you shouldn't take their words at face value.

Douthat: Much, much like a superintelligent A.I.

Kokotajlo: Sure. But in terms of — I can at least say that the sorts of things that we've just been talking about have been discussed internally at the highest level of these companies for years.

For example, according to some of the emails that surfaced in the recent court cases with OpenAI, Ilya, Sam, Greg and Elon were all arguing about who gets to control the company. And at least the claim was that they founded the company because they didn't want there to be an A.G.I. dictatorship under Demis Hassabis, who was the leader of DeepMind. So they've been discussing this whole dictatorship possibility for a decade or so at least.

Similarly, for the loss of control — you know, "what if we can't control the A.I.s?" — there've been many, many, many discussions about this internally there. I don't know what they really think, but these considerations are not at all new to them.

Douthat: And to what extent — again, speculating, generalizing, whatever else — does it go a bit beyond just, they are potentially hoping to be extremely empowered by the age of superintelligence? And does it enter into, they're expecting the human race to be superseded?

Kokotajlo: I think they're definitely expecting the human race to be superseded.

Douthat: But superseded in a way where that's a good thing. That's desirable, that we are encouraging the evolutionary future to happen. And by the way, maybe some of these people — their minds, their consciousness, whatever else — could be brought along for the ride.

You mentioned Sam Altman, obviously one of the leading figures in A.I. He wrote a blog post in 2017 called "The Merge," which is, as the title

suggests, basically about imagining a future where human beings, or some human beings — Sam Altman, right? — figure out a way to participate in the new super race. How common is that kind of perspective — whether we apply it to Altman or not — in the A.I. world, would you say?

Kokotajlo: So the specific idea of merging with A.I.s, I would say, is not particularly common. But the idea that we're going to build superintelligences that are better than humans at everything, and then they're going to basically run the whole show and the humans will just sit back and sip margaritas and enjoy the fruits of all the robot-created wealth — that idea is extremely common. I think that's what they're building towards.

Part of why I left OpenAI is that I just don't think the company is dispositionally on track to make the right decisions that it would need to make to address the two risks that we just talked about. So I think that we're not on track to have figured out how to actually control superintelligences, and we're not on track to have figured out how to make it democratic control instead of just a crazy possible dictatorship.

Douthat: I think that seems plausible, but my sense is that it's a bit more than people expecting to sit back and sip margaritas and enjoy the fruits of robot labor. Even if people aren't all in for some kind of man-machine merge. I definitely get the sense that people think it's speciesist, let's say ——

Kokotajlo: Some people do. Yeah.

Douthat: To care too much about the survival of the human race. It's like, OK, worst case scenario, human beings don't exist anymore. But good news, we've created a superintelligence that could colonize the whole galaxy. I definitely get the sense that people think that way.

Kokotajlo: There are definitely people who think that. Yeah, yeah.

Douthat: OK, good. Yeah, that's good to know.

Kokotajlo: [Chuckles.]

Douthat: So let's do a little bit of pressure testing in my limited, limited way of some of the assumptions underlying this kind of scenario — not just

the timeline but, whether it happens in 2027 or 2037, the larger scenario of a kind of superintelligence takeover.

Let's start with the limitation on A.I. that most people are familiar with right now, which gets called hallucination. It's the tendency of A.I. to simply seem to make things up in response to queries. You were earlier talking about this in terms of lying and outright deception. I think a lot of people experience this as the A.I. making mistakes, and that it doesn't recognize it's making mistakes because it doesn't have the level of awareness required to do that. A recent story in The Times reported that in the latest publicly available models — which you've suggested are probably pretty close to cutting-edge — there seem to be trade-offs where the model might be better at math or physics, but guess what? It's hallucinating a lot more.

Are hallucinations just a subset of the kind of deception that you're worried about? When I'm being optimistic, I read a story like that and I'm like, OK, maybe there are just more trade-offs in the push to the frontier of superintelligence than we think, and this will be a limiting factor on how far this could go. But what do you think?

Kokotajlo: Great question. First of all, lies are a subset of hallucinations, not the other way around. I think quite a lot of hallucinations — arguably the vast majority of them — are just mistakes, as you said. So I use the word lies specifically. I was referring to specifically when we have evidence that the A.I. knew that it was false and still said it anyway.

But also, to your broader point, I think that the path from here to superintelligence is not at all going to be a smooth, straight line. There's going to be obstacles to overcome along the way. I think one of the obstacles that I'm actually quite excited to think more about is what you might call reward hacking. In "AI 2027," we talk about this gap between what you're actually reinforcing and what you want to happen — what goals you want the A.I. to learn — and we talk about how as a result of that gap you end up with A.I.s that are misaligned and that, like, aren't actually honest with you, for example. Well, excitingly, that's already happening. That means that the companies still have a couple of years to work on the problem and try to fix it.

One thing that I'm excited to think about and to track and follow very closely is: What fixes are they going to come up with? And are those fixes going to actually solve the underlying problem and get training methods that reliably get the right goals into A.I. systems, even as those A.I. systems are

smarter than us? Or are those fixes going to temporarily patch or cover up the problem instead of fixing it? That's the big question that we should all be thinking about over the next few years.

Douthat: Well, and it yields a question I've thought about a lot as someone who follows the politics of regulation pretty closely. My sense is always that human beings are just really bad at regulating against problems that we haven't experienced in some big, profound way. You can have as many papers and arguments as you want about speculative problems that we should regulate against, and the political system just isn't going to do it.

In an odd way, if you want the slowdown, if you want regulation and limits on A.I., then maybe you should be rooting for a scenario where some version of hallucination happens and causes a disaster, where it's not that the A.I. is misaligned, but — this sounds sinister — it's that it makes a mistake and a lot of people die somehow because the A.I. system has been put in charge of some important safety protocol or something, and people are horrified and say, OK, we have to regulate this thing.

Kokotajlo: I certainly hesitate to say that I hope that disasters happen and people die, but ——

Douthat: We're not saying that. We're speculating.

Kokotajlo: I do agree that humanity is much better at regulating against problems that have already happened when we learn from harsh experience. Part of why the

Kokotajlo: I do agree that humanity is much better at regulating against problems that have already happened when we learn from harsh experience. Part of why the situation that we're in is so scary is that for this particular problem, by the time it's already happened, it's too late.

Smaller versions of it can happen, though. For example, the stuff that we're currently experiencing: We're catching our A.I.s lying, and we're pretty sure they knew that the thing they were saying was false. We're pretty sure it was a blatant lie despite the fact that that wasn't what their instructions were and that wasn't what their training was supposed to train them to do.

That's actually quite good, because that's a small-scale example of the thing that we're worried about happening in the future, and hopefully we can try to fix it. It's not the example that's going to energize the government

to regulate because no one's dying. It's just a chatbot lying to a user about some link or something.

Douthat: And then they put it in their term paper and get caught.

Kokotajlo: Right. But from a scientific perspective, it's good that this is already happening because it gives us a couple of years to try to find a thorough, lasting fix to it. And I wish we had more time, but that's the name of the game.

Douthat: OK. So now two big philosophical questions, maybe connected to one another. There's a tendency, I think, for people in A.I. research making the kind of forecast you're making to move back and forth on the question of consciousness. Are these superintelligent A.I.s conscious and self-aware in the ways that human beings are? I've had conversations where A.I. researchers and people will say: Well, no, they're not, and it doesn't matter because you can have an A.I. program working toward a goal, and it doesn't matter if they are self-reflective.

ut then, again and again, in the way that people end up talking about these things, they slip into the language of consciousness. So I'm curious: Do you think consciousness matters in mapping out these future scenarios? Is the expectation of most A.I. researchers that we don't know what consciousness is but it's an emergent property, and if we build things that act like they're conscious, they'll probably be conscious? Where does consciousness fit into this?

Kokotajlo: This is a question for philosophers, not A.I. researchers — but I happen to be trained as a philosopher.

Douthat: Well, no, it is a question for both. Since the A.I. researchers are the ones building the agents, they probably should have some thoughts on whether it matters or not if the agents are self-aware.

Kokotajlo: Sure. I think I would say we could distinguish three things. There's the behavior: Are they talking like they're conscious? Are they pursuing goals? Do they behave as if they have goals and preferences? Do they behave as if they're experiencing things and then reacting to those experiences?

Douthat: Right, and they're going to hit that benchmark.

Kokotajlo: Definitely, yeah.

Douthat: Absolutely, people will think that the superintelligent A.I. is conscious. People will believe that.

Advertisement

Kokotajlo: Because in the philosophical discourse, when we talk about: Are shrimp conscious? Are fish conscious? What about dogs? Typically what people do is they point to capabilities and behaviors, like, look, a dog can recognize itself in a mirror. It seems to feel pain in a similar way to how humans feel pain and has these aversive behaviors, and so forth.

Most of that will be true of these future superintelligent A.I.s. They will be acting autonomously in the world, reacting to all this information coming in, making strategies and plans and thinking about how best to achieve their goals. In terms of raw capabilities and behaviors, they will check all the boxes, basically.

There's a separate philosophical question of, well, if they have all the right behaviors and capabilities, does that mean that they have true qualia? Did they actually have the real experience, as opposed to merely the appearance of having the real experience?

That's the thing that I think is a philosophical question. I think most philosophers, though, would say, yeah, probably they do, because probably consciousness is something that arises out of this information processing cognitive structures. If the A.I.s have those structures, then probably they also have consciousness.

However, this is controversial, like everything in philosophy.

Douthat: Right, and I don't expect A.I. researchers to resolve that particular question. It's more that on a couple of levels, it seems like consciousness as we experience it, as an ability to stand outside your own processing would be very helpful to an A.I. that wanted to take over the world.

So at the level of hallucinations, if they produce the wrong answer to a question, the A.I. can't stand outside its own answer-generating process in the way it seems like we can. If it could, maybe that makes the hallucination process go away. And then when it comes to the ultimate worst-case scenario that you're speculating about, it seems to me that an A.I. that is conscious is more likely to develop some kind of independent view of its own cosmic destiny that yields a world where it wipes out human beings than an A.I. that is just pursuing research for research's sake.

But maybe you don't think so. What do you think?

Kokotajlo: So the view of consciousness that you were just talking about is a view by which consciousness has physical effects in the real world. It's something that you need in order to have this reflection, and it's something that also influences how you think about your place in the world.

would say if that's what consciousness is, then probably these A.I.s are going to have it. Why? Because the companies are going to train them to be really good at all of these tasks, and you can't be really good at all these tasks if you aren't able to reflect on how you might be wrong about stuff.

So in the course of getting really good at all the tasks, they will therefore learn to reflect on how they might be wrong about stuff. If that's what consciousness is, then that means they'll have consciousness.

Douthat: OK. That does depend, though, in the end, on a kind of emergence theory of consciousness like the one you suggested earlier. Basically, we aren't going to figure out exactly how consciousness emerges, but it is nonetheless going to happen.

Kokotajlo: Totally. An important thing that everyone needs to know is that these systems are trained; they're not built. So we don't actually have to understand how they work — and we don't — in order for them to work.

Douthat: OK. So then from consciousness to intelligence, all of the scenarios that you spin out depend on the assumption that, to a certain degree, there's nothing that a sufficiently capable intelligence couldn't do.

I think a lot hinges on this question of what is available to intelligence. Because if the A.I. is slightly better at getting you to buy a Coca-Cola than the average advertising agency, that's impressive, but it doesn't let you exert total control over a democratic polity.

Kokotajlo: I completely agree. And so that's why I say you have to go on a case-by-case basis and ask: OK, assuming that the A.I. is better than the best humans at X, how much real-world power would that translate to? What affordances would that translate to? And that's the thinking that we did when we wrote "AI 2027."

We thought about historic examples of humans converting their economies and changing their factories to wartime production. And we asked: How fast

can humans do it when they really try? Superintelligence will be better than the best humans, so they'll be able to go somewhat faster.

And so maybe, instead of in World War II, when the United States was able to convert a bunch of car factories into bomber factories over the course of a couple of years, well, maybe then that means in less than a year, maybe six months, we could convert existing car factories into fancy new robot factories, producing fancy new robots.

Douthat: But if we're looking for hope, this is a strange way of talking about this technology. We're saying the limitations are the reason for hope.

Earlier we talked about robot plumbers as an example of the key moment when things will get real for people. Then it's not just in your laptop. It's in your kitchen and so on. But actually fixing a toilet is, on one hand, a very hard task. On the other hand, it's a task that lots and lots of human beings are quite optimized for.

I can imagine a world where the robot plumber is never that much better than the ordinary plumber. In that world, people might rather have the ordinary plumber around for all kinds of very human reasons.

And that could generalize to a number of areas of human life where the advantage of the A.I., while real on some dimensions, is limited in ways that at the very least — and this I actually do believe — dramatically slows its uptake by ordinary human beings.

For instance, right now, just personally, as someone who writes a newspaper column and does research for that column, I can concede that top-of-the-line A.I. models might be better than a human assistant right now by some dimensions. But I'm still going to hire a human assistant because I'm a stubborn human being who doesn't just want to work with A.I. models.

To me, that seems like a force that could actually slow this along multiple dimensions if the A.I. isn't immediately 200 percent better.

Kokotajlo: So I would just say this is hard to predict, but our current guess is that things will go about as fast as we depict in "AI 2027." They could be faster, they could be slower, and that is indeed quite scary. Another thing I would say is that we'll find out how fast things go when the time comes.

Douthat: Very, very, very soon.

Kokotajlo: The other thing I was going to say is that politically speaking, I don't think it matters that much if you think it might take five years

instead of one year, for example, to transform the economy and build the new self-sustaining robot economy managed by superintelligences.

That's not that helpful if the entire five years there has still been this political coalition between the White House and the superintelligences and the corporations, and the superintelligences have been saying all the right things to make the White House and the corporations feel like everything's going great for them, but actually they've been deceiving them.

In that scenario, it's like, great, now we have five years to turn the situation around instead of one year. That's, I guess, better. But how would you turn the situation around?

Douthat: Well, let's end there.

In a world where what you predict happens and the world doesn't end — we figure out how to manage the A.I. and it doesn't kill us, but the world is forever changed — and human work is no longer particularly important, what do you think is the purpose of humanity in that kind of world? How do you imagine educating your children in that kind of world and telling them what their adult life is for?

Kokotajlo: It's a tough question. Here are some thoughts off the top of my head, but I don't stand by them nearly as much as I would stand by the other things I've said, because it's not where I've spent most of my time thinking.

First of all, I think that if we go to superintelligence and beyond, then economic productivity is just no longer the name of the game when it comes to raising kids. They won't really be participating in the economy in anything like the normal sense. It'll be more like just a series of video-game-like things that people will do for fun rather than because they need to get money — if people are around at all. In that scenario, I guess what still matters is that my kids are good people, and that they have wisdom and virtue and things like that. So I will do my best to try to teach them those things because those things are good in themselves, rather than good for getting jobs.

In terms of the purpose of humanity, I don't know. What would you say the purpose of humanity is now?

Douthat: Well, I have a religious answer to that question, but we can save that for a future conversation.

I think the world that I want to believe in, where some version of this technological breakthrough happens, is a world where human beings main-

tain some kind of mastery over the technology, enabling us to do things like colonize other worlds. To have a kind of adventure beyond the level of material scarcity.

As a political conservative, I have my share of disagreements with the particular vision of "Star Trek" — but "Star Trek" does take place in a world that has conquered scarcity. There is an A.I. computer on the starship Enterprise. You can have anything you want in the restaurant because presumably the A.I. invented the machine that generates any food you want.

So, if I'm trying to think about the purpose of humanity, it might be to explore strange new worlds to boldly go where no man has gone before.

Kokotajlo: Oh yeah. I'm a huge fan of expanding into space. I think that would be a great idea. And in general, also solving all the world's problems, like poverty and disease and torture and wars. I think if we get through the initial phase with superintelligence, then obviously, the first thing to do is to solve all those problems and make some sort of utopia, and then to bring that utopia to the stars would be the thing to do.

The thing is that it would be the A.I.s doing it, not us. In terms of actually doing the designing and the planning and the strategizing and so forth, we would only be messing things up if we tried to do it ourselves.

So you could say it's still humanity in some sense doing all those things, but it's important to note that it's more like the A.I.s are doing it, and they're doing it because the humans told them to.

Douthat: Well, Daniel **Kokotajlo**, thank you so much. And I will see you on the front lines of the Butlerian Jihad soon enough.

Kokotajlo: Hopefully not. I hope I'm very wrong.

Douthat: All right. Thanks so much.

Kokotajlo: Thank you.