So let me start this one by situating us in time.
So we taped this podcast just a few hours after Open AI released GPT-4, which is their most powerful AI language model yet.
The thing that is remarkable here that I really keep trying to push people to key into is how quickly these systems are getting better, how steep the curve of improvement is.
We talk about some benchmarks here on the show, but GPT-4 now passes a bar exam in the 90th percentile, the 90th percentile.
Let's sit with that.
We were nowhere near fully understanding what the last generation of these models could do.
Now we're on to the next one.
And the same is probably going to be true here.
GPT-4, it is extremely powerful.
It is at times a-nerving.
We barely understand what it can do or how to use it.
And yet everybody is racing ahead.
GPT-4 also is going to be a toy, a diversion, compared to what we're going to have in three or five years.
If you heard the commentary I released on AI on the podcast over the weekend, you heard that what I'm most focused on is how society manages to adapt to the speed at which these systems are improving, or else slows them down so we have more time to adapt, to understand them, to think about how we want to integrate them.
And one piece of feedback I got in that commentary was, what are you really saying we should do aside from be anxious?
And I don't just want people to be anxious.
So this conversation gets more into specifics, both on what these systems can do now, why they could be so disruptive, where they seem to be going, and what we can and I think should do, or at least the kinds of things we should do, in response.
Kelsey Piper is a senior writer at Vox, where she's been doing great coverage for years on AI and the technologies and the culture and the safety and economic questions around it.
She joins me now.
As always, my email is reclineshow at nytimes.com.
Kelsey Piper, welcome to the show.
Thanks.
I'm glad to be here.
So one of the reasons I wanted to have you on here is you've done a lot of reporting.
You spent a lot of time with folks inside the AI world.
So you're a little bit more in this hot house than a lot of us.
And when you talk to them over the past couple of months, what do they say?
What are they seeing or what are they thinking about that from the outside would not be obvious?
Yeah.
So I think not just in the last six months, since like at least 2019, you had people saying, this is going to overnight, not today, but in 10 years, overnight transform our world.

We are going to be able to automate any work that can be done remotely.
We are going to be able to bring us the next 100 years of drug discovery and development in the space of maybe a couple of months, maybe a single research project.
We have this extremely powerful thing we're on the brink of.
And in 2019, when you heard that, I think you had to be a little skeptical.
Like this is Silicon Valley.
We're not new to the hype cycle.
Sure you're going to change the world.
And now I think you're still hearing the same things, mostly from the same people, but you have to take them a little more seriously.
They are in fact releasing things that can do things that everybody thought were impossible.
They are developing the newest models as fast as they can.
And AI says, we think the way for humanity to do well on AI is to raise a head as fast as possible.
This is in a statement a couple of weeks ago.
This morning, GPT-4, only a couple of months behind chat GPT, which hit 100 million users like overnight.
It's not even that they're saying new things, it's that it's harder not to take them seriously.
So when you say overnight, this idea that sometime in the future, all of a sudden, the history of AI will speed up.
That we're making progress fast, but it's a couple of years, a couple of months between major new systems, and then you get these ideas of quantum leaps really quickly.
What are people talking about there?
What are they looking at?
So I think the most concrete version of this is that right now all of the work done on improving AI systems is being done by humans, and it's a pretty small number of humans.
And of course, one of the jobs that they are trying to automate is their job, the job of making better AI systems.
And when they succeeded that, if they succeeded that, then you can instead of having hundreds of people working on building better AI systems, you can have thousands, you can have a million, because you can just make a copy of an AI system.
That's like its fundamental economic potential is that you can make a copy of it.
You can make an AI system that is as good at writing as you, and then you can have a thousand of them writing about all different topics.
You can make an AI system that's a really good programmer, and then you can have all of the programmers working with or possibly replaced by that thing.
So things are already moving fast.
And I don't think the case for like taking AI really seriously crucially depends on they'll start moving even faster.
Like the current pace is fast enough to me to be disastrous.
But if you ask like what's my gut instinct about what's going to happen, it's that at the point where we figure out how to make AI researchers out of AIs, then things will move real fast.
And that sounds very sci-fi, as a lot of things do, because we live in a sci-fi world increasingly.

But one of the things that I know is striking to a lot of people in this world was when GPT-3 really proved able to code, even though it hadn't really been trained to do so.
And so then as people began moving more towards optimizing AI systems to create code, and you expect we'll do even more of that, train in on even more code, get even better at that.
Once AIs can code, which it can already do, when you just assume for the improvement curve, then the idea that you would have an AI that codes AI, that's not a giant leap.
That's simply finishing the sentence in a slightly different way.
Yeah.
I guess a lot of it is just there is so much commercial incentive to make an AI system that can code.
Programmers are really expensive.
They're scarce.
It's hard to hire them.
If you have an AI system that can do most of what they can do, companies would be shelling out huge amounts of money for it.
And that means companies are shelling out huge amounts of money on developing it.
And this is just also true, more true of the specific skill of building large AI systems.
So I don't think you need this things will accelerate premise necessarily to take AI really seriously.
But my best guess is that there are people right now trying to develop AI systems that can develop AI systems, because there's people trying to develop AI systems that can do every human job.
Well, let's talk a bit about how fast it's going now.
So this is a little bit by coincidence, but the morning we're talking open-air release GPT-4.
And they did not, according to them, train GPT-4 specifically to pass any tests.
But like GPT-3 and 3.5, it absorbs a lot of language from the internet.
It begins to create statistical correlations for lack of a more precise way of saying it between terms.
And out of that emerges a kind of thing that seems to act in the real world like understanding.
So they released this table where they showed the performance of GPT-3.5 and GPT-4 on a battery of tests that human beings find pretty hard.
So when you try to give GPT-3.5 the bar exam, it is in the 10th percentile of people.
When you give it to GPT-4, it's in the 90th percentile.
GPT-4 passes the LSAT at the 88th percentile.
GPT-3.5 was at the 40th percentile.
I found this one kind of funny, but also interesting that it passes the advanced Somelier theory test.
That's for wine in the 77th percentile, GPT-3.5 was in the 46th percentile.
So these are fairly general tests of knowledge and reasoning.
And obviously there's information about them out there.
I mean, these have essay questions, et cetera.
Now we're in an area where the improvement is very fast and you can actually match it against human beings on things that I would not have said five or 10 years ago.

Oh, that's just a structured data quiz.
There's an old argument that we say artificial intelligence is anything that a system can't do yet and machine learning is anything it can already do.
But how do you think a bit about this debate over truly intelligent versus stochastic parrots as people like to call it versus sort of a very advanced autocomplete?
So it sort of calls into question what intelligence is, right?
What is going on in our brains when we think?
I think there is a sense in which it's just a glorified autocorrect, but I think that's also like a sense in which most of the time we are also just a glorified autocorrect.
Like what humans do is we read lots of things, we talk to lots of people, and then we put it together into our own sort of framing that articulates something that we wanted to articulate.
And it turns out that most of that you can just speed something, the entire internet, and then make it really, really good at pattern recognition and then prompt it and it does that.
And I guess you can say, well, that's not intelligence.
But if you're going to declare, well, that's not intelligence, I think you are kind of in a position where you're also declaring that about humans.
And I don't know, what do you even get out of saying?
Humans aren't really intelligent.
We still did build civilization.
We still did build nuclear bombs.
If we're not intelligent, then I'm not sure intelligence is the thing we care about here.
Well, what people say to give credit to this view is that they don't understand.
Like we wouldn't call it calculator intelligent even though it gets the right answer if you try to multiply 37 by 25, and frankly, most human beings can't do 37 by 25 in their mind.
And so there's something about the quality of understanding, of knowing why you're saying what you're saying.
And I kind of agree that on some level, these things don't know why they're saying what they're saying.
But on the other hand, I think one of the demos that keeps coming up that unnerves me the most is when the systems get asked to explain why a weird joke is funny and then they do.
The difference between knowing why you're saying what you're saying and being able to give me a count of why you're saying what you're saying that is convincing to me seems like it's getting very thin.
Yeah.
Like maybe you can say, well, it could tell you a human who said the thing I just said would probably have been having this thought process.
And if we want, we can say, but that's not its thought process.
Its thought process is this incomprehensible alien thing with no resemblance to the human thought process it just gave you.
I think that's true.
I don't find it very reassuring, right?
Like whatever it's doing may be different from the account it can give, but it does have the skill to produce the account still.

And I don't know that we can say that that incomprehensible thing that it's doing isn't really reasoning.
I think that gets to one of the things that I actually find the most unnerving, which is that the more inhuman the statistical correlation process happening at the center of these systems becomes, the more human the experience of interacting with the system becomes.
And that fundamental loss of legibility between we have some mental model.
When I ask you a question, I have a rough, not a perfect mental model of what's happening as you think about the answer.
But as we make these systems more complex and they have more billions of connections and they have all these layers and parameters and nodes, and if you tried to get it to print out what it's doing, when you ask it why a joke is funny, you'd get like a billion calculations that are happening that mean nothing to you.
And so when I think forward 15 or 20 years, I think that's what unnerves me, like a complete sort of loss of human comprehension of a lot of the interactions we're having and decisions being made around us.
And not just loss of our comprehension, but our capacity to have comprehension over it.
This is some of the more scary and sort of depressing visions of the future you see from like Holden Karnofsky or from Ajay Akatra or like a world where
Poor AI researchers.
Yeah, AI researchers at the Opa Philanthropy Projects.
They're thinking about the future and where this is taking us.
And a description they gave that really stuck with me was decisions by companies about where to build new factories, about what products to invest in, about what kind of R&D to do, are made by AIs.
The AIs explain their answers to us perfectly.
They write beautiful memos, but we can no longer tell if the memos that they're writing and the decisions that they're making are towards goals that we have, towards things that we want, or if we've built some self-sustaining inhuman thing that is, you know, sort of doing its own thing now.
Because we can ask and it will give us a beautiful answer and we just have no way any more to tell if that beautiful answer is right or wrong.
And is that, I mean, one thing you could say, because you're getting here at a problem of deception, right?
When you say that, I think there, or maybe there are a couple versions of it, one version of it that I think you're pointing towards is we tell an AI to do something.
We have mis-specified what that thing is.
And so we think it's doing what we want it to do and it's actually doing what it wants to do and trying to trick us along the way.
But could he say it's true for human beings?
I mean, in businesses, aren't people constantly putting a factory somewhere and actually it turns out they got a kickback?
It wasn't because they really wanted the factory there for economic reasons?
Yeah, I think that some degree of this thing has its own goals is like accounted for in the system.

We are used to working with people who have their own goals.
Nothing really relies on everybody being totally loyal to us because that doesn't work.
I do think that the more alien something is, the less that works because some of that works on laws and anti-bribery rules and oversight, but some of it works on the fact that very few people are going to build a factory that pumps chemicals into the atmosphere that poison everybody alive.
They just wouldn't do that.
And an AI might do that if it had a reason to do that.
You can't just count on the fact that it's probably basically a reasonable person and not an omnicidal maniac the way you can with a human because humans are all from the same share of values, of priorities, of kinds of thoughts.
Like you said, you can sort of guess what's going on inside my head.
And even if you're getting similar behavior out of something that's radically alien, you don't know that its scope of action is the same, that it won't do things that you would never think of doing.
So I want to go back to the pace of improvement.
And really, this is one of those conversations where I don't want to start shouting into the microphone, but the thing I really want to emphasize is that I'm not sure most of us have any sense of how exponential the curve of improvement both is and is becoming.
So when you talk to the people working on these systems, how would you describe what they think will be true about them that isn't now in five years?
And how do you describe that in 20, noting that neither of those values are all that long?
Yeah.
These are like before our kids grow up, tears, things.
So I think that a lot of people anticipate we are not very far from systems that can do any job you can do remotely, anything you can do on your computer.
So coding, like secretarial work, receptionist work, journalism, everything like that.
I think that some of those same people would say, we can invent better scientists and do science a lot faster and come up with, you know, whatever discoveries humanity would have plotted along to over the next hundred years, whatever those we would have learned from studies and research, getting it very, very fast from AI systems that can do that work sort of overnight or in the space of days, just running much, much faster.
And then I think some of those people are like, and then we get, you know, a fascinating cool Star Trek future where material scarcity is over, no one needs to work, no one has to work, and all of the goods of all of this AI work they've done are distributed to all humankind.
And some people are like, if we did a poor job of managing this, of regulating this, of having oversight of this process, that is where, you know, we die or we get some other pretty catastrophic outcomes from effectively unleashing a civilization of aliens that we don't have a good way to work with or trade with or understand.
So I want to back up a little bit from the sort of utopia or dystopia, just because, like, honestly, I think you lose people here.
And I don't even know what I think of those, because for instance, on the scientific one,

I know there's this big thing among the AI community that it's going to create a scientific super explosion, but most science that I can think of requires you to run a bunch of experiments in the real world.
And I'm not sure exactly how much faster an AI really makes it to laboriously do phase three trials of vaccines on like human beings or, I mean, I can imagine ways it could speed things up, but I do wonder a little bit about the potential for an overestimation of analytical intelligence.
But let me go back to something you were saying a minute ago, because I think it's closer at hand.
When you say people believe, you know, five, 10 years, I think this is a fair description.
Many people in this sector believe that within five to 10 years, you will have systems that can do any work you can do remotely that is not in the physical world.
At least as well as like the median person doing that work right now.
That's a pretty profound both.
I think we think about that as a possible economic change.
It's also a kind of metaphysical change for humanity.
Tell me just a bit about why people believe that to be true.
So you don't need to get all that much better than GPT-4 plus some careful training and some careful shaping of the models that we have and some additional work on reliability and stuff like that to get something, you know, it can pass the LSAT.
It can also do a lot of what lawyers tell me is the sort of menial first few years of law work of putting contracts together and reviewing them.
It can program.
It can program quite well.
It is not all that much improvement required before it can program better than the average person you might otherwise hire to program.
It can write, I think my writing is better than it's, but I think my writing is better than it's by like a much smaller margin than a year ago or than two years ago.
And so at some point you have to be like, probably next year it will be better at writing than me or if not next year, probably in a couple of years.
And I think some places, probably the New York Times will still employ human journalists, but I think a lot of places have already sort of tried to see if they can get away with automating huge amounts of work that used to be done by humans, art too.
Used to be that humans would draw art and now for many, many of the purposes that we used to hire humans for AI art is where it's at.
And I think that you're going to see a ton of that.
Think about this with like weaving.
When we invented the automated textile loom, some people still bought like hand woven rugs because they really liked the, you know, high end hand woven thing.
And you wouldn't say that automated rugs are like better than hand woven rugs in every way.
Every rug in the world is an automated rug and almost all rug work is done by machines.
And I don't think very many people want to be like a niche cottage industry for people who prefer human labor while nearly all of our labor is better done by AI.

To think about writing for a minute, because a lot of jobs rely on writing.
True for journalists, true for lawyers, true for social media marketers, true for people who do publicity, true for just a huge amount of the modern, you know, what gets called the knowledge economy.
And one sort of metaphor that I've thought about in my own head for thinking about it in journalism is you're going to move from the highly valued skill set, the probably more valued skill set being that of the writer to being that of the editor, the person who looks over a piece of work, tries to think about, is this true or not?
Did this miss something obvious?
Does the structure of this argument make sense?
But just putting aside like what you think of those skill sets, just a truth that anybody who's in a newsroom knows is that there is a small proportion of editors to writers.
An editor will often supervise somewhere between, depending on the organization, four and eight or nine writers.
And that's true, I think in a lot of areas, like go to social media marketing, which I know a little bit about, you will have a number of junior marketers who are overseen by one more senior person.
You go to law.
You have a number of younger lawyers who are overseen by a partner, et cetera.
And on the one hand, so you have a skill shift towards management, but also oversight just by its nature.
You are overseeing a number of agents, for lack of a better term here.
And as you were saying earlier, the key thing here is these things are cheap to copy.
And so just when you think about just that, right?
If it's just true, they're like, okay, editors, like we're not getting rid of all skills and writing.
Do we need more editors and fewer writers?
Well, we only need fewer editors and we need writers.
So it gets really quickly into land that will be pretty disruptive for people.
And even if for some reason, or somehow we passed laws, or just turns out that it hallucinates a little bit too often to make this possible, the scale people want to imagine, I still think it's going to shake a lot of people to feel all of a sudden that skills they spent their whole life generating may not matter that much.
I mean, I think it's unnerving.
When I put something in, I'm like, write about polarization in the style of Ezra Klein.
It's not there yet, but it is what you would get if you had a ninth grader who'd read me for a while, write about polarization in the style of Ezra Klein.
And if it gets to where like me five years ago, it's going to shake me a little bit.
Yeah.
And we never know.
It could all peter out next year.
But my bet is that that point where it's you five years ago is not actually all that far away because it's a ninth grader now and it was the sixth grader a year ago.
And it was the fourth grader a year before that.

And I think that just the magnitude of the economic disruptions like call centers do a ton of customer service work.
A lot of that is probably going to be automated in the next few years.
That's like a ton of high paying jobs in poorer countries where people speak English lost in places where those people don't necessarily have nearly as good alternative jobs.
And maybe some of these like massive collapses of various industries will be replaced by new industries created by AI.
But I don't think there's like a guarantee in general economics that your industry being wiped out doesn't just leave you notably worse off.
It can just do that.
If you had specialized skills and then they're no longer valuable, it can be that you would just have lower earnings for the rest of your life.
It typically does mean that.
I mean, we know this from the economic literature.
People sometimes pretend that economies adjust over time.
But when somebody's factory gets closed, they do not typically get rehired for a higher wage.
This is not how it goes.
I think that's something we were talking about this before the podcast started and I want to bring it in here because I think it's emotionally important.
I was saying a minute ago that I've read for years the arguments about whether I kills us all and I just even if I can follow their logic, it is very hard for me to emotionally connect to them.
Something in the past couple of months, the thing that has kind of messed with my head though was thinking about my own children and not whether there would be a world for them to grow up in people, but just what kind of world, how different it might be for mine.
Because as I was looking at this, like I think for people who have seniority in the world, there are a lot of reasons as I can come for you immediately.
And it will for some people, obviously.
And I do think that even if the technology moves quickly, societies adapt more slowly, there are regulations and rules and oversight and so on.
But it just seems, if you look at this economically, if you look at just the possibility of people being surrounded by inorganic, non-human things that feel like intelligences to them, sort of the world of the movie Her, which I was rewatching recently and I cannot recommend watching it enough at this moment, but there's a company called Replica that is explicitly trying to do this in a great piece in New York magazine recently about people falling in love with their replicants.
Those are going to get way better.
I tried that out.
My replicant was super boring, maybe I'd get better over time, but it won't be boring for long.
And if you read this sort of Kevin Roos transcript with Sidney Bing, whatever, like that wasn't whatever else you want to say about it.
That was an interesting, like diverting conversation.

I just think we're on the, like for my kids, like in my kids, not just life, but like growing up period, we're on the cusp of a world that for them may be profoundly different than the world I was prepared to raise them for.
And something about that has been the crucial line I stepped over where I went from being like, this is interesting and I like to read about it to holy shit.
Yeah.
No, I think that there's a sea change and it is going to mean that they grew up in a world that in some ways we can barely recognize.
And I think, I don't know, I am in favor of progress and change and technology and I mostly like Silicon Valley.
I have all these friends here who have all these crazy hyped startups and I like, I appreciate that worldview in many ways, but I do feel this sense of ideally you have the frontiers pushed forward.
We see some gaps that it's leaving some awful things that are resulting, we adjust, we regulate, we figure out where the problems are, we invented a banking system and that was great.
We realized the banking system had all these catastrophic failure modes and we try and learn how to patch them and how to manage them and how to keep failures manageable.
We invent nuclear weapons and that's terrifying, but we also over time develop institutions surrounding non-proliferation and norms and stuff like that.
We came closer to a nuclear war in the early period after nukes were developed than recently and kind of like if we have enough time, we can make almost anything that happens to us a good thing.
And the problem is if we have enough time, I don't think that that process can magically happen overnight.
Let's talk about why this might not happen.
So what do you think is the best argument for why this hits a wall in your term?
We get to something like GPT-5, it's a bit better than what we have now.
But we've seen, I mean, driverless cars were supposed to be all over by now and you can drive around and see little Waymo cars, but they've all got people in them and they've not been able to get to the reliability they really needed.
We didn't replace all the radiologists, like that was supposed to happen.
IBM Watson's got functionally sold off her parts.
A bunch of these predictions have come over the past, it's called 15 years and they didn't quite pan out and there was often a moment where people thought it was inevitable and then inevitability turned to disappointment at least over this time period.
So if we plateau not far from where we are now, why do you think that will have been?
Yeah, so my pessimist voice goes, okay, these systems are trained on all of the data on the internet and they are not going to be any smarter or any better or more discerning than everything on the internet.
There's some arguments that they could be because you need like higher quality to replicate it than just is necessarily present in it, but you could certainly say, no, they're just kind of going to be as good at making stuff as everything on the internet.
And the average quality of everything on the internet is not outrageously high.
It's not like super impressive.

It's good at like writing things that seem to flow for a minute, but there's often, and I'm not talking about like AIs here, there's often not a lot of like underlying thought there.
There's often not a lot of accounting for complexity.
So maybe the best thing you get out of training and really large language model on the internet is something that can in like tiny fragments imitate a human who thought about it, but without any of the underlying cognition.
And so maybe that means that those things never actually, they can pass a standardized test, they can pass a standardized test amazingly well, we have to rethink standardized tests.
They can write a college essay, but that really reflects the fact that college students are not putting actual cognition and thought into their essays.
They can imitate maybe the sort of low level journalism that's just write this event that happened, but they can't imitate synthesis.
They can't say anything original or distinctive.
You could totally imagine that it would not be very surprising.
And does that change the world?
I don't know.
I feel like it makes the world weirder, but it doesn't bring in some of these scenarios that are really transformative, I think if it turns out that there is some degree of actual cognition that doesn't go into most human utterances, but is important for human work to be of economic value, then maybe it's not there.
But when I say that, then I'm like, I sound like all the people who are like, it needs reflectivity, it needs the ability to consider counterfactuals, it's not going to have those.
And every time I've heard someone say, oh yeah, yeah, you can't just feed it the internet and get reflectivity, you can't just feed it the internet and get logic, you can't read it the internet and get coding ability.
It turns out you can feed it the internet and get those things.
So I guess what I say to my pessimist voice is like, yeah, maybe.
I find that view very compelling, though.
And I mean, I've had Gary Marcus on the show who holds, I think, a version of that.
And I probably am like 30% where he is.
But the place where I maybe feel like I differ from people who hold that view is that I think making something as good as the internet could be more important.
And again, to what you're saying, I'm going to go weirder.
Like I really, really focus in on the word weird here.
Like my big common AI last year was all about the word weird.
My recent significant common AI was all about the word weird.
I think that can make something very weird.
I think it can make something economically quite important.
If you could make something as good, I mean, if you can do a good job, because one thing about the big models is like GPT-4 isn't trained to be a lawyer, it's just trained on the internet.
But if you do the fine tuning in the area of making it a really good lawyer, it's going to be way better than GPT-4, even at its current level of capability at being a lawyer.
So I could see how you can make things as good as a pretty good lawyer, as good as a

pretty good journalist, as good as a...
It is a little hard for me to understand how you get this superhuman intelligence training something on human data, both because it's functionally, I think at some level averaging out human data, and also because, and this does go to the point where understanding might be important.
It doesn't really have a great way.
It seems to me of separating what it has read that is true and false.
I think about something like the masking debate about whether or not masks work for COVID.
And one thing within that is it, if you just inhaled every study on masking, putting without any really profound priors about how to think about any of that, I don't know what you would get, but it isn't obvious to me the mechanism by which you would get something way better than what we have, because you wouldn't really have the capacity to run a new study yourself.
If we could just run all the randomized controlled trials of masking we wanted, and maybe we should be doing that, we would know a lot more.
And also, how would you know what studies to believe and do you just end up in hawk to whatever people have done the most of?
I actually find that view very persuasive, so I'm curious what you don't find persuasive in it.
So I guess that might happen.
I feel so much uncertainty about AI, and a lot of what I feel is that if you lay out anything kind of plausible sounding, I'm like, yeah, that might happen.
I don't feel like we can rule that out.
I also don't feel like we can rule out the crazy extinction scenarios.
I feel like ruling any of those out would require a better grasp on what it is we are doing than we have.
So I'm like, that might happen.
But my gut is that it won't.
And if I try and dig into why not, I'm like, to imitate what we would say in a situation, you could say that takes less intelligence than it took us to say it.
But I think in some cases, it might take more, like you have to in some ways be able to model a person in order to come up with what that person would say.
And there are people who have lived our lives on the internet for the next 25 years, and you would think that just being fed that person would give you a pretty good internal picture.
And to build a generator of everything as your client has to say, just a perfect predictor, something that just writes what you have to say, something probably has to have whatever processes you have, or processes that successfully imitate whatever processes you have.
And to the extent that your processes are truth tracking, to the extent that the thing you're doing involves like noticing which sources seem reliable and reasoning in your head, like step by step, how would masks work?
How would a mask study fail to tell us about masks?
You can already get GPT-4 to do that.
It will give better analysis of mask studies if you tell it.
Reason step by step.
How good is this study?

What are some things that might confound to this study?
And so I'm kind of like, maybe it can do that prompted, but it can't put it together and it can't use its synthesis of our reasoning to do its reasoning, even just to imitate us.
But it probably can.
It probably just, in order to write something as good as an Azure client column, which is something we are trying to train it to do, it will probably have to pick up a lot of the actual underlying process and not just the output of that process.
This is one of the things I find most, again, just weird interacting with the systems, but it's now a highly replicated result.
If you say, tell GPT-4, or GPT-3.5, which is what I actually know, give me the answer to this, as if you were really smart.
It will give you a smarter answer.
You mentioned reason step by step, right?
A bunch of ways you can do a prompt, right?
You've done versions of this, like, tell me this, as if you have an evil other side and you're trying to hide it from me, which is to say that if you give it the instruction to model a kind of person, it can model that person, again, not really trained to do that at any level.
I mean, you can ask it to write in the style of me and it will, which was, I read it, it was a little embarrassing.
I thought it did catch some things, like, yeah, I do do that, a lot of m dashes, and that's weird to me.
And it's funny.
It's something that ends up happening in her, which comes from 2013.
I mean, it's a minute ago.
But the AIs themselves begin creating models of people from the past they want to be able to interact with.
So just one thing that could just make the future weirder, even before you get to ideas of superhuman intelligence or anything else, anyone with a really profound written corpus that you can feed tens of thousands, hundreds of thousands of their words into a learning program that's already been trained on the whole internet and then tell it to particularly seem like this person, you might just be able to give people all kinds of fairly capable, mimicked human beings from the past and present.
I don't exactly know what to say about that, but it seems strange to me.
And then if you do that, then like something loose on physics or whatever, I don't know they get way beyond where humans are, but it can probably get way beyond where most humans are, which is already enough, I think, to destabilize quite a bit of our sense of ourselves ontologically.
Yeah, even in that sort of world where they don't get any smarter than us, there is just still some really radical, terrifying transformations from the ability to create human intelligences sort of at will and customize them for almost any purpose that normal humans can do and run them for as long as you want and stop running them when you want.
You could imagine like every student having specialized teachers designed to work with

that student customized to have the ideal personality for working with them.
You could imagine a lot of us have AI friends or AI lovers who are more attentive and don't have any needs of their own, and you could imagine in some sector that outcompeting normal human friendship.
Personal assistants are obviously a big area where most of us can't afford a personal assistant but would love to be able to say like, hey, can you schedule a doctor's appointment for me or have someone who proactively says, hey, you haven't been to the doctor in three years, you should really get on that.
And I think all of that is coming even if none of the stuff about building on that for greater intelligence happens at all.
When I talk to the people working on these systems, I'll often ask, what can you make it do that I can't?
Because I think something people don't always realize is that even the demos we're being given are highly constrained in what they can do.
And something I've heard from a few sources is that you can already tell these systems to accomplish something in the real world that you can do through the internet, but that has multi-steps like, you know, it needs to figure out how to rent a car and do things like that.
And it can.
It can come up with that plan and execute something in the real world.
And sometimes it will come up with a plan that is not the kind you would think of or not the kind you would want it to come up with, but it knows enough.
It's not just that it can give you an answer, but it can execute an action.
And that's already, but we don't know it.
I mean, this is actually one of the things that worries me the most really about policy debate and sort of where we are on this, that even before you take into account the steepness of the improvement curve, what they already have is quite a bit beyond what we know.
We only really have access to open AI and open AI derivatives like Bing.
We don't have access to what Google has built, the public.
We don't really have access to what Meta has built.
We don't really know a bunch of the systems being built in China.
And we don't have access in any of these cases to complete open systems where they haven't been highly constrained for our use.
And here I'm talking in we, the public, or any kind of sort of mass sign up.
The one moment I think anybody got access to a system that wasn't highly well constrained, which was like the Bing Sydney debacle, it really occasioned a moment of reflection and many cases I think fear.
And that's just something going on here in the background that I just like don't quite know how to convey, but we don't even know what we're dealing with now exactly.
And there's a lot that's not yet released, even though places are doing a pretty fast release cycle, there's just inevitably part of an engineering processes that you're ahead internally of where you're going to go externally.
So I actually think GPT-4, which is now out, is pretty close to the state of the art.
I don't think there's like tons of more sophisticated models anyone has.

It's hard to get side-by-side comparisons, but where I have been able to get them, GPT-4 is very good.
But what there is is like applications that are pretty obvious that people have set up and have working, yeah, stuff like an integration of chat GPT or GPT-4 with TaskRabbit or something like that so that it can just send instructions, pay a person to follow those instructions and get stuff done in the physical world, totally already done.
Similar with Voice, well, Google had this years ago and kind of backed off of it, but Voice that sounds like a person and can call and make an appointment for you and sound totally natural on the phone and is an AI assistant making appointments for you.
So when people are like, so if AIs wanted to do bad things, they're a good thing, they're stuck on computers, right?
And it's like, I mean, we built them tons and tons of tools with which they can directly interact with the physical world in many, many ways and we're going to keep doing that.
New York Times games make me feel like I'm amazing.
Wordle makes me feel things that I don't feel from anyone else.
The Times Crossword puzzle is a companion that I've had longer than anyone outside of my immediate family.
UNKLE?
Is that he's welcome?
Cool.
No.
You should know what it's called.
Okay.
I started Wordle 194 days ago and I haven't missed a day.
I absolutely love spelling B. I always have to get G'd in, yes.
I really like words that use few letters but give you a lot of points, palapa, falafel.
I've seen you yell at it and say that.
That should be a word.
Totally should be a word.
My proudest crossword achievement is my four-minute, 54-second Saturday.
Crossword heads, you're going to be impressed by that.
When I can finish a hard puzzle without pins, I feel like the smartest person in the world.
When I have to look up a clue to help me, I'm learning something new.
It gives me joy every single day.
Join us and play All New York Times games at nytimes.com slash games.
Tell me about what people call the alignment problem.
So we haven't yet had to grapple all that much with the fact that we don't know if these systems have goals and we don't know if they want things and we don't know if their actions are trying to accomplish things.
And we don't super need to know that with existing systems because if you prompt TPT4 to write an article in the style of Ezra Klein or whatever, then it writes the article in the style of Ezra Klein.
And it's not doing the open-ended tasks where it might matter what its goals are and it probably doesn't have goals.

I don't like making any certain statements about it guys, but it probably doesn't have goals.
But when we build these systems, we want them to have certain goals.
We want them to be helpful, not say racist things, not embarrass the companies that release them.
We want them to be truthful.
We want them to do what we prompted them to do.
We don't know how to tell them any of that.
We're just like doing it all with reinforcement.
Like whatever it does, something we can say good or bad and maybe with more gradations in there, like how good or how bad, but that's what we've got.
We're trying to use that very blunt lever of reinforcement to point it at our goals.
And depending exactly how you do this, there's a bunch of ways that what you teach it to do is probably not what you actually want it to do.
You want it to be truthful, but you might be wrong in some case, and so you reinforce it poorly for a true answer because you thought that answer was false.
And so it learns, oh, don't be truthful with the humans.
Figure out what the humans believe and tell them what they want to hear, which is different.
You were trying to teach it truth, but you were going from your own understanding.
And it might notice that if it has a little personality, if it's sympathetic, if it's nice or for that matter, if it's a little bit sassy, that gets better reactions than if it's sort of bland and boring.
So maybe it develops and displays a bit of a personality because it gets reinforced, and that's what it thinks we want to know.
And you can get a fair bit of deceptive behavior this way, right?
If the thing that we are trying to teach isn't the thing that we actually want, and it is incentivized to give us what we want to see rather than actually be on our side or whatever.
How much of that is already happening?
Could I have a question about deceptive behavior?
Good question.
So the problem is that it's hard to tell, right?
We know that the models say false things all the time.
We know that they say false things when they know what the truth is.
We can easily, you know, I can just go to chat, TPT, get it to say something false, and then ask in a different context, and it'll be like, yeah, that answer was false.
We know that.
Is that deception?
Probably not.
Probably it wasn't like trying to trick me.
It was just, you know, giving the answer it thought I wanted.
But it makes it very hard to tell when there is intentional deception with intent of like accomplishing some goal that the AI has.
And this is extra hard when we don't know if the AI's have goals.
And some people will say they clearly don't have goals, because how would something like

that have goals?
I'm just kind of like, I don't know, evolution was just like selecting repeatedly on ability to have babies.
And here we are.
We have goals.
Like, why does that process get you things that have goals?
I don't know.
They just showed up along the way.
Language language models are just like selecting repeatedly on like token prediction and then reinforcement.
Does that have goals?
I think if anybody tells you like yes or no, they are overstating what we know and what we can know.
One of the questions here that I find interesting is this question of coherence.
Can you talk a bit about that?
The thing I think of when I think about coherence is to what extent does it make sense to say that I have goals?
I like have lots of wants.
Some of them contradict each other.
I want to get ready for my interview and I also want to lie in bed another 10 minutes.
I want to stay up late playing video games and I also want to be a functional adult who doesn't stay up late playing video games, you know?
And humans act coherent and goal directed in some contexts, like we have priorities and we will predictably do the things that accomplish our priorities.
But in other contexts, we don't.
And you could imagine that if AI systems have goals that they could be like that.
We don't know if they have goals.
We don't know how to tell what their goals are.
We kind of know how to make them work towards our goals, but only when the work is something that we can verify and we're eventually going to want work from them that we can't verify or like can't verify at the relevant scale.
And that does dramatically change the safety landscape.
But I don't know, if we knew that evolved aliens were coming to Earth in the next 10 to 20 years and we knew that being evolved beings like us, they were incoherent and they had lots of competing goals and impulses.
And you know, some of those were to turn Earth into resources that they could use to build more factories, but some were to keep us in zoos and some were to like leave us a nature preserve and some were to like invite us to participate in their geopolitics and some were to ally with us against these other aliens they're in a war with or whatever.
I'd be scared.
I wouldn't be sure we would all die, but I would be nervous.
That would feel like a bad situation, even though they're incoherent just like we are and even though they might in principle be willing to work with us.
One of the things you talked about in one of your pieces is what people call specification

gaming.
If you imagine an AI that, I don't know, however we built the alignment, we can actually communicate
with and it can kind of understand us and it does care about trying to figure out what we want.
I think that's one world, but we're going to have a lot of things that we really can't communicate with on that level and in the near term, specification gaming seems really worrisome, particularly if AI has become a big part of hiring processes or as we've already seen, predictive policing and so talk a bit about specification gaming.
So say you do give your AI an objective that you're trying to have it maximize, which we do sometimes when we're like building game playing AI's or whatever, or yeah, if you're doing a hiring process, you might train an AI where your training is aimed at the maximum share of the resumes that the AI passes through, in fact, result in a job offer.
We want this AI to be really good at filtering and correctly predicting who gets a job offer or whatever.
AI systems will be very creative sometimes.
That's what some of the examples of specification gaming we have demonstrate.
Satisfying that constraint we gave them in a really out of the box way we would never have thought of.
If you're having them play a game and they win by getting the most points, they will do weird stuff like find a glitch in the game, they will repeatedly race in circles forever because they can get more points that way.
They will trigger some kind of bug in the code that just causes them to end up with a really big number that happened once.
I find that kind of freaky.
The AI just figured out how to take some random action that caused the number to be really big and we don't really know how it noticed that.
It may have stumbled on it randomly.
Who knows?
So we are building these AI systems and giving them these narrow objectives and I don't think this is how the world ends or anything, but it's certainly a dynamic that, yeah, we're facing all the time where if the hiring manager who looks at the resumes is racist, then the AI will maximize its success probability by like never passing along a resume of the wrong race because then that one would be less likely to get hired.
That's like an alignment failure with society.
The AI is like enabling a bad actor to do bad things, but there's also ones where the AI is like not working for anyone.
It's just accomplishing its objective in a way that would be unacceptable, cost wise to everybody, but we don't have adequate oversight to notice that that's what it's doing.
Like if you had a stock trading AI, there are so many horrifying things that you could get up to that like modify stock prices, right?
You can trigger wars to modify stock prices.
You can hack your competitors to modify stock prices.
You can trigger financial collapses.

And so I do think that even though this isn't the apocalypse scenario or whatever, whenever you're putting these like creative deeply in human algorithms in charge of making some metric go up, you would hope we would be sensible enough to make really sure that that was under lots of oversight and that it didn't have the power to do things in weird ways.
And that's not what happening because that's never what's happening in AI.
You can say like, oh, well, that sounds bad, but it wouldn't be a problem if we just took these like three very reasonable measures to counter it.
But we're not taking those three reasonable measures like ever.
There's always someone who is just doing the naive foolish version of the thing.
And as the systems get more powerful, the failures get more powerful.
I actually think the stock trading example is something weirdly that I worry about a lot because if you think about what sector of society right now has functionally unlimited money to pump in building the absolute best or licensing the absolute best AI systems that you can, well, hedge funds, other kinds of quant trading operations.
Then if you think, well, also separately, what segment of American or global commerce is most used to using algorithms, it itself does not understand completely recklessly in an effort to get even the most narrow edge over competitor, you'd say, again, high speed and algorithmic trading.
And I mean, I can kind of keep going down like if you might say like, what's a good industry that's really good at wielding political influence so people don't get in its way because it's really good at hiring an excellent lobbyist to high speed, but what is an industry where if you actually had a really, really effective algorithm that people didn't care that much about oversight because they were trying to get it to maximize a particular metric in this case, returns might be able to cause a lot of damage to a lot of people.
You say, oh yeah, something around high speed and algorithmic trading.
And I don't exactly know what we're planning to do about that, except as you say nothing because like the current theory of what we're going to do here is nothing, I think.
Which is mostly what we've done with high speed and algorithmic trading so far.
That's a little bit unfair, but I think that has gone way beyond where there's any social benefit to having algorithms competing with each other to do things nobody understands to exploit the small differences and prices at a given moment.
But that's just a near term thing that seems very concerning to me.
You don't need to have any kind of particularly far term views about like how AI will develop to just think a lot of damage could happen by unleashing powerful systems nobody understands in a place where they're trying to justify something that people really want maximized and may not think as they often don't think through the consequences of maximizing it.
Yep.
I think there almost certainly will be disasters in this space and I just hope that they're at a scale we can learn from instead of at a truly catastrophic scale.
And I also think there's a general principle here of we are creating these alien intelligences that are creative, that think of problem-solving solutions we didn't think of, that do weird things that no human in their position would ever do because it's a route to their goal.
People who play against them in chess or go talk about this, how sometimes the algorithm will make a move that just no human would ever make and they don't see until 20 steps

down the line how brilliant it was.
And I worry that in the world of try and make this company's stock price go up, that tendency towards genuine creativity will take the form of, oh, I bet that if I cause an outage at this power plant, then stocks will move in this predictable way that I can make a lot of money off of.
And if you don't have oversight, stuff like that happens.
Or if you don't even know how to have oversight.
Yes, because you can't see what it's thinking because it's not thinking in any way we understand.
This gets to something within the culture of this community that I want to talk about.
You cite, and I've cited in pieces too, this really unnerving survey of machine learning researchers from last summer, from 2022, it asked people basically what they thought the chances were that working on AI, if they got to high-level machine intelligence, the consequences for humanity would be really, really, really bad, like human extinction bad or some other complete displacement of humanity.
And the median response was they thought there was a 10% chance of that happening.
The people working on this said, yeah, 10% this thing that I am in some way or another trying to bring into being or trying to worry about how it comes into being, I think 10% chance humanity is completely pushed off the board in a fundamental way.
And you could say that's just a survey, but I've talked to enough of the people at high levels in these companies to know that that survey is not that weird.
Yes, my impression also is that that survey, it might be something of an overstatement, but the position there is not rare.
There are plenty of people who will tell you, oh yeah, I think this is it.
I don't know exactly how to communicate to people, but I've spent enough time in this community, and you've spent more than me, that these systems are being designed by people who themselves do not believe they really know what they're going to do.
Just out that my column is talking about it as a kind of summoning, and I've also had that thought that this is like people opening a portal, and they will freely tell you that they hope a demon doesn't crawl through, but they can't guarantee a demon doesn't crawl through.
And even if you don't believe they're right, like you believe this is crazy, or maybe it's hype, they believe this.
And that actually worries me a lot because you talk about value alignments.
I think a lot of people might say, if the thing you're working on has a 10% chance of destroying humanity, don't, or go really, really slow.
Some people do believe that, but there is a lot of now competitive pressure to speed up to be the first one.
There's pressure between US and China, so geopolitical pressure, Google does not want to lose to Microsoft, it does not want to lose to Meta.
Can you just talk a bit about how you understand the culture here, and the psychology of working continuously to bring into existence something that you know you don't understand, and where you think the consequences have a really, really high chance of being unacceptably bad?
Yeah.
So I want to second what you were just saying about how difficult this is to convey, because

you have these conversations, and you get this like somewhat terrifying sense of that people understand themselves to be making extraordinarily high stakes decisions, decisions that I think the average person on the street would say, that should absolutely be the prerogative of a government with a lot of careful consultation, maybe some kind of international framework, and you have people who are like, we're going to do it, we're going to do it ourselves, we thought about it, and we think it's going to go best if we do it ourselves, and we do it now.
And it's like very hard to step outside of that bubble and simultaneously say, these risks seem real and not sci-fi and scary.
And also, you know, whatever you think about that, the people who are doing this do think that there's something real here.
So there's like, why are they doing it?
And I think some of it's just the love of science, right?
Like there's, I've been reading The Making of the Atomic Bomb, and a recurring like theme in that.
Oh yeah, why are you reading that?
I, you know, have two young kids at home and another one on the way, I never have time to read, and so I was way behind on my book list, and then I got the invite to come on your show, and I was like, this is my excuse to read all of the books that I'd been wanting to read, but now it's for work, I have to read them, so I have something to recommend on Ezra's podcast.
And that is how I finally got some time to read The Making of the Atomic Bomb.
And one of the recurring things in it is, it's a very personality driven compared to, I guess, what I was imagining before I picked it up, like who were these men and what were they thinking?
And it does feel like there's a kind of person who just once they realize it can be done, there's an overwhelming urge to do it.
There's just a, one of the points that the author makes at the beginning of the book is that we might dream of a world where physicists said, on principle, we will not build an atom bomb, but we were never close to that world.
Physicists would build an atom bomb as soon as it occurred to them that you could.
You know, and there were some very brave and very notable people who tried saying like, this seems really bad, what should we do about this?
But overall, if you want to stop dangerous technologies that has to be on a structural level with like governance and regulation and international agreements and voluntary between lab agreements, you can't count on people, unfortunately, to just say, this is ludicrously dangerous, we're not going to do it.
Because I guess some share of people are just like, this is ludicrously dangerous and I'm going to do it.
I guess this is like the classic tale of hubris in some ways, like the oldest known flaw in human nature is you see the tree of knowledge and you're like, yeah, I'm going for it.
Well, it's kind of a selection effect too.
Yes.
Which is to say that if you were somebody who thought this was such a bad idea that you

definitely wouldn't do it.
And you don't do it.
And by definition, you're not the person doing it.
And there's a lot of incentive for people to believe the arguments that they should do it.
Yes.
Right.
But there are a lot of people who work in AI safety and who are around this community and spend all the time thinking about reasons maybe not to do it or reasons to try to make it, you know, slow it down or how to make it go more safely.
But in the end, they're not the ones who run the companies who get the investment that gives them the compute power to do it.
There are so many people, very good machine learning researchers who I know who nobly made the decision to not play this game because they thought they might kill us all.
And so they are out of this game and they are, you know, working in academia on much smaller models or working on regulatory frameworks, they're doing good work.
But they're not the people at these labs because they said no, and they walked away.
And so we have the people who, for whatever reasons, didn't walk away.
And so I think within those, you can cut a couple groups, right?
So Jan Lacoon, who's an important AI pioneer and he's one of the key people at META, I believe.
He's their chief AI scientist.
He just thinks it's all dumb and that an AI has no interest in taking over or killing us all or whatever.
So it's just not going to.
And this has all gotten very farfetched.
Then you have people who run a couple of the other labs, maybe Sam Altman at OpenAI would fit into this framework, folks over at Anthropec in a different way, I think, are more careful than OpenAI.
And who believe that AI could do great good and do believe that there's a chance of really bad things happening and think that by working on the models and experimenting on them, that you make it both likely or you understand the bad things so you can prevent them and also maybe trigger society coming in and saying, okay, here's how we're going to regulate this early enough.
And I don't think it's a crazy view.
I mean, I do think that there are more conversations happening about how to think about and regulate
AI because OpenAI has been releasing models into public use that Google and Meta are not.
So I don't think it's a crazy view, but they are all now in a race.
And I worry that the race dynamic is just a lot stronger than the caution dynamics.
And this is something that I know you've been writing about and thinking about.
So can you talk a bit about this question of the race because something I know has been worried about in AI for a long time, but I think it's undeniable now that it's happened.
Yes.

So there's a couple of different things that drive getting new powerful models, right?
One is how much money tech companies are willing to sink in and how much talent they have working
on it.
And race dynamics often mean that these companies are willing to dump a ton more money on this because they want to stay ahead of their competitors.
Microsoft is willing to spend billions and billions of dollars for a shot at catching
Google in search.
Google is presumably willing to spend billions and billions of dollars defending its spot
in search for the same reason.
And so you get bigger models faster.
And that's like one thing you can analyze, like maybe it's good to have these bigger models faster under these race conditions because then we're here talking about them.
We probably wouldn't be here talking about them if this was still all under the hood
because it would be a little too crazy.
But the other thing is that when you're going fast, safety becomes a liability and things that you could do to make your models less dangerous or to check more thoroughly if your models are dangerous, work you could do on like mitigating risks and even just understanding them goes out the window because you're trying to beat someone else and you want it in commercial production as fast as possible.
We would be way safer against AI takeover scenarios, really scary catastrophes.
If AI systems didn't have open access to the internet and the ability to contact anybody
they wanted and do anything they wanted on the internet.
But we've given them that.
Of course, we've given them that because it was like slightly more profitable than not giving them that and so as soon as you open that door for everybody to be competing with each other, then the people who are careful and cautious are slower than the people who are careless and incautious and the people who are the most willing to throw safety considerations out the window are the ones who are getting their biggest models out fast.
And that puts us in a situation where even if it turns out to be totally possible to make AI systems safe, like there's a very straightforward technical solution to alignment.
We know what exactly what we need to do.
We still have a problem because somebody can just decide that would be a little expensive.
That would delay our launch by two weeks.
That would require us to hire another engineer like Y-Botler.
So when I search for any AI app that I might want to check out in the app store, one thing I always notice is that there's the app I was looking for and then there's a bunch of fast follower apps right beneath it.
And my sense is that the kind of sub premium AI models a couple years ago were just total crap.
Yes.
But now you're looking at things that are not coming from open AI or Microsoft or Meta or Google, but they're as good as things were a couple of years ago, which is like now not that bad.

And then in a couple of years, they're going to be as good as the things are now.
And if you are open AI, you have a really, really high reason because you have so many awesome ways to make money here to make sure you're restricting use of your model to only things that are going to get you good press or something like it.
But you know, you have AI's that it now seems you can train them on somebody's local patterns pretty easily and it can then call my mom and because I've got a lot of my voice out there in the world.
You could train down on my podcast and it could call my mom and it could sound like me having been kidnapped and needing a ransom.
This has been done.
So, okay.
So it's already happening.
And these are getting better faster than you actually don't.
For things that are just kind of bad, you don't need the best system in the world.
You can increasingly just use kind of crappy system and the crappy systems are like the best systems in the world from a couple of years ago.
That's just going to keep happening.
And that seems, again, like something we don't really have a framework for thinking about.
We don't really know what to do about.
But I wonder if you could talk a bit about that, that the opening up not of the state of the art, but of the sub state of the art to functionally like every scammer, advertiser, just anybody who wants to make a buck in the world.
And the kind of CD underbelly of AI, in the same way that happened to crypto, though maybe that was like the overbelly or something because it kind of started there.
But the CD underbelly of AI strikes me as a place where a lot of damage is going to be done really soon if it's not already being done.
Yeah.
And this is again, something where I'm like, if things were moving a little slower, I'd be like, yeah, there's some scary stuff and we'll adapt to it and people will learn not to trust calls from their kidnapped sons and, you know, we will get better at spam filters to combat spammers and we will, you know, get better at making comment sections on the internet functional.
I do get the sense in like the last year that spam is ahead of spam detection in terms of like-
Yeah, really seems to, something seems to have changed.
Yeah.
I feel like we went through this golden period where spam detection was better than spam and you mostly didn't encounter spam all the time and it is over tragically.
And now I get a lot of spam on Twitter, in my email, on Facebook and I'm like, I guess we are back in the domain where spam is better than spam detection.
And if you have a lot enough time, then I think it's possible for people to develop the social technology and the literal technology to counter that kind of thing.
But when everything's moving really fast, you know, people learn how to address one thing and then we get the AI that can like carry on a long convincing conversation with

your mom and convince her to send her bank details in the course of like this very long conversation that also references a bunch of details from your childhood that you mentioned on the podcast once.
That's harder, you know?
How do we get to that point where it's easy to cause so much damage and so expensive to detect?
And again, I'm just like, I wish we were moving a little slower.
I wish we had more time to like figure this stuff out.
Well, you say that like it's impossible.
What's the actual difficulty?
I mean, we have made a lot of things slower.
We can't clone humans and we don't even though it is a possible thing to do and I'm sure there are scientists who would like to clone humans.
We've actually been quite good at maintaining an international ban on cloning human beings.
There's all kinds of biological weapons research that we just like don't really seem to do anymore.
We did it for a while and we actually used them and then we have kind of non-proliferation packs that I'm not saying they are perfectly effective, but they are effective.
And you can come up with a bunch of different things like that.
There's certain forms of genetic modification we just don't do.
Katya Grace, I thought had a who's an AI researcher at a really good piece on this about how the AI community has a little bit of this weird tendency to say, look, maybe it'd be good if we slow down, but it's completely impossible to ever coordinate people such that they slow down on anything.
And her point, which I think is true, is that that's actually not true.
We coordinate people to not do things or slow down on things all the time.
In some ways, it seems to me the biggest barrier here is actually the people who just want to rush forward on AI are matched by the people who don't think that rushing forward on AI matters.
So it's not that much interest in slowing anybody down because, what, it's like toys on the internet.
But I don't know.
I do think there's a problem here where the AI community seems to have convinced itself that it's so hard to get people to do anything that you couldn't possibly coordinate them around this.
I don't see why that would be true.
Yeah.
So I think we can and should slow down.
It's very hard when things are moving this fast to get a good feel for the possible.
You could imagine a fairly sweeping set of government regulations in the next few years that set a cap on how large you can make a model unless it is subject to some form of democratic oversight and extensive testing for safety, where we check things like, can this send a convincing phishing email that persuades 1% of users to run code on their computer?

If it can, you can't generally release it yet.
Can this model, like, give detailed instructions on how to carry out a terrorist attack that if followed would be successful?
If it can, you can't release it yet.
There's all kinds of scary things that are not very far off.
And we could just draw a card cap and be like, look, if the model in an audit of some kind with dedicated testers whose job is to elicit this behavior, if the model can do the dangerous thing, no, you cannot deploy that model.
Or you're liable if you deploy the model, which maybe amounts to the same thing while we're talking about big companies.
We could do that.
We could do that tomorrow.
I think there's plenty of people in Washington who are thinking about something like this.
And I think what's possible, you're so right that it's a social question.
It's a question of what everybody else believes.
If everybody wakes up and goes, we want to have more steering power over the world our kids grow up in, we could just do this.
There are like very reasonable rules you could put in place that would significantly change the pace of this work.
And I hope that happens.
I'm not pessimistic.
I've heard from enough people that they want something to change, that they're scared, that they believe it can and should go differently than it's going, that I'm like, that's what it takes and maybe we can make something happen.
But I am kind of the default thing that happens is that the regulators are years behind the ball.
We have to do a very dedicated, very concerted effort to set up a system where regulators are on the ball.
It's a hard problem.
It's a hard problem I think we should be working on, but it's a hard problem.
Tell me a bit more of that list, because I do think one thing that makes problems harder is when the suite of possible things you might do about them feels opaque.
So a lot of people don't understand that much about it, and they really don't understand what you might do to regulate it without completely decapping it.
Are you just saying you can't research language models anymore?
I mean, that seems crazy, and I think would actually be kind of crazy.
So is there a rough consensus among the safety people you talk to about these are the three or four things we really wish would happen?
I would say there is an in-progress effort to build that consensus.
I think conversations I heard this year that I did not hear last year included like, what are some clear-cut benchmarks like send a phishing email that persuades 1% of people who receive it to run code on their computer.
That's a pretty clear specific.
It's obvious to anybody who's thought about policy a little bit why that's scary behavior.

It's pretty easy to use an auditing process to determine whether a model has that capability.
You'd have to send some phishing emails, so I guess there's ethics concerns.
But I feel like a regulatory framework can work with that kind of thing.
So that's conversation I didn't hear last year that I am hearing this year that gives me some optimism, because I think that's what it would look like to have regulators be on the ball instead of behind it.
I do think that people who are working on this effort are coming at it from all kinds of different places.
Some of them are like, well, what I really want is for all of this to stop.
I can't get that.
So I will try and set some benchmarks that have an achievable chance here.
Some of them are like, I really want to be doing my amazing AI research, but I want these safety people off my back.
So I'll agree to some standards that I can meet that don't get in my way too much.
And a lot of people have different threat models and different pictures of which AI capabilities are going to be the scariest ones in the near future or in the medium future.
So it's a difficult process to build that consensus.
And I think we're still in the early stages of it.
What are a couple other specifics that you think are interesting?
They don't need to be consensus, but you just think they are provocative.
So if an AI system can, via emailing instructions to a human assistant, get a copy of itself spun up on Amazon Web Services independent from its creators and so that another copy of it can run on AWS.
That seems like a benchmark.
You can define that behavior in an audit.
You can see whether the model can in fact do it successfully.
And if it can, then that's like an indicator of some of the really scary scenarios where models decide to operate independent of human oversight because that lets them achieve their goals better.
So that's maybe a benchmark.
If your model has the capability, when prompted, to give a human instructions on how to copy that model on, run it on AWS independently, that's a scary model.
We don't want you to widely deploy that model.
When you say when prompted here, one interesting thing to me about that example is that I feel like there's a set of behaviors or capabilities where like we are worried to humanely use you this way.
Then there's this kind of further off ideas that like we think you might act this way, you being the model.
And in that world, if we think the model develops situational awareness of itself and its world, such that it's then being prompted by some idiot human like, hey, do you mind emailing and copy of yourself to the Amazon Web Services to replicate in the world where you think that the AI has become deceptive, how well does that actually work?
Having this auditing process is very tough, right, like it's a very technical, difficult work to figure out.

How would we elicit this behavior if it is possible?
But I think that in general, it is easier for a model to display a skill when it's elicited by humans than unsolicited.
So we will probably successfully identify models that are capable of giving a human instructions on how to copy them to a new AWS server sooner than we will get models that can do that and think of hiding it and successfully hide it from us.
Now that's a gamble.
Maybe we do our tests and we're like, oh, it's good, these models can't do it.
And in fact, these models have passed the threshold where they can do it, but suspect it's not in their interest.
But that feels less likely to me.
That requires the model to develop two skills at the same time, like the deceiving us and the capability to survive and spread itself or whatever.
I feel a lot more optimistic if we're consistently checking, hey, can you do this?
And then halting things and calling a red flag once it can do this, because I think that will probably be before it develops the ability to do that deceptively without our knowledge or awareness.
So let me make a critique of the AI safety community here a bit, which is, I don't want to say people are too focused on the absolute worst that can happen.
That's a good thing for some people to be focused on.
But I find that there's an unbelievable lack of focus on what seems to me to be an incredibly important near-term question, which is the business models behind these models.
And I think in general, the sort of community this comes out of is a little unreflective about capitalism, right, effective altruists generally and rationalists and so on.
But these models are going to develop really fast along the pathways that make them a profit.
Like at some point, the stuff has to begin turning around money.
And the models that develop the most, they're going to turn around the most money.
And I think something most of us believe about the internet at this point is it would have been better, all things considered, if surveillance advertising had not become the dominant model for all of the biggest players.
I said before on the show, I worry about AIs of this magnitude being hooked up to surveillance advertising.
But there's other things I also worry about them being hooked up to.
And I think people are really unreflective at this point or under focused on this question of regulating how you can make money within machine learning is actually very important for affecting the path of machine learning in the long run.
And I'm not sure that I know.
In fact, I'm sure I don't know what is safe and what is not safe here.
But a world where there's billions of dollars to be made because the US government had set up a series of prize competitions for things like the protein folding problem, which Deep Mind was able to solve, you'd get AI systems that could do one thing.
In a world where there's billions of dollars to be made and trying to get me to buy things and manipulate my behavior, you'll get AIs to do another.
I really worry about the most obvious profit models being a spur then to create AIs.

If you create AIs and the main way they make money is manipulating human behavior, you've created the exact kind of AI I think you should fear the most.
But that's the most obvious business model for all of these companies.
It's what they already do.
Whereas that's not what, say, Pfizer does, but Pfizer's the one creating the AI.
This seems obvious to me and I can't seem to get anybody to care about it.
No, I'm with you.
I guess I would be very happy to see regulation that was aimed at how you can make money off of your powerful AI systems to sort of start and discourage building manipulative powerful systems because I do think manipulative powerful systems are how we get into very serious trouble.
So I think there's a ton of blind spots in every AI community just because being part of a subculture that the rest of the world didn't take seriously until two months ago is weird and distorting to people and because the kind of people who cared about AI when it was weird are weird and because there's some ways in which it's like a little unhealthy to spend all your time thinking about these futuristic sci-fi scenarios.
And it's not that they're so far in the future that they don't merit thinking about.
But I think you have to be very grounded and I think a lot of people get a little ungrounded.
And I think a lot of the stuff you're talking about, you're right and it's very grounded.
It's just like this will go according to the logic of what makes money unless we successfully make rules about how you can make money off of it.
That seems true.
And if you've spent all your time trying to think about this as a technical problem, like how do you make the AIs not lie?
How do you notice if they're lying?
Then you maybe just get a little bit of tunnel vision and you don't see other opportunities to substantially curtail scary stuff that are outside the range of stuff you're thinking about.
So there's a kind of dimension of competition between companies that are functioning within a couple of miles into each other or a couple of states of each other.
And then I find that if you push on that conversation for a while, the next space of fadles and becomes geopolitical competition, that of course we have to go as fast as we can because if we don't, China will.
And even if somebody could come to an agreement with China, like you just wait long enough and Russia will or Qatar will eventually or somebody will whose values you don't share.
And so that there is in all of these dimensions of competition, a view that there's such a huge return to being the first to retain AI dominance, that if you simply don't somebody else will, and by virtue of you being the one who didn't, you will like their values less.
So I'm curious how you first hear about the geopolitical dimension discussed within this world and second, what you actually think of it.
Yeah.
So I think you describe it completely correctly.
People will talk about Microsoft versus Google and if you're like, but we could regulate, we could slow down, then the thing that comes out is sure, but then China.

I think this is kind of silly.
First of all, international cooperation is not impossible.
For many technologies there it is possible, especially ones that are very dangerous.
It is possible, especially once we well establish what about them is dangerous to arrange international
controls of some kind.
To the extent you think AI is something that could be bad for everyone, then I think you should also be a little optimistic that as we develop good benchmarks for identifying AI being bad for everyone, we can also work internationally on it.
So if you think that this is the highest stakes thing in the world, then I think rushing ahead and doing it poorly and doing it without buy-in and while a lot of people who work closely in a research community with you are saying, you are going to kill us all.
You are going to catastrophically destabilize the world we live in because you're worried that some other country is going to get there first.
It just feels to me like you've got to be making some kind of logical mistake there.
I come back to the making of the atomic bomb, which as you said, it's pretty clear why it was on my reading list.
This process where the terror that Nazi Germany was building a bomb was a big motivator to start the bomb project.
Then when it came out that not only was Nazi Germany not building a bomb, they'd never been close and there was no chance they would get the bomb.
In fact, we were going to win the war with them without the bomb.
One scientist quit the atomic bomb project when he learned that actually there was no Nazi bomb project and there was no international competition.
Nobody else did.
Partly because moths to a flame can't stop summoning the demon.
The ones they know they can, they have to find out if they will, but partly because I suspect that for a lot of people, they wanted to find a reason.
They found that reason.
If that reason wasn't applicable, find a different reason.
Also, China's good at espionage.
From a pure boring geopolitical, let's leave all the ethics out, the easiest way for China to get really powerful models is to steal the really powerful models that we are making.
I think it would slow them down a bunch if we were not making really powerful models that they could then steal.
You ask these companies, oh, you're so worried about competition with China, you must be taking really serious security measures then because you don't want your model stolen by China.
What kind of steps do you take to make sure that a nation state cannot steal your stuff?
I'm like, oh, we're a startup.
I just can't take it that seriously.
I don't like the CCP, the Chinese Communist Party.
I don't want them to shape the future of human civilization with extremely powerful AI systems.
I just do not believe that the best way to prevent that outcome is for us to, you know,

with this shoddy security and this unclear sense of what we're going to do with the powerful models, zoom straight ahead.
I don't buy it.
I think all that is true and I actually really, I think that point about espionage is really well taken and this is where my background as a politics person, I actually worry because my read is the most powerful stakeholder on AI in the federal government is a national security apparatus that traditionally they have done the most work on AI.
They have the most freedom to work on AI and they've just thought about it more.
I think that most of the energy on AI until very, very recently is coming from them and they are inclined to look at the world in this way.
That's one thing that concerns me a little bit in terms of how it gets thought of politically.
Even if you look at some of the recent investments in AI, in the Chips and Science Act, that's very explicitly framed as an Anti-China Act.
That was where I think the biggest recent play on this came from.
I mean, there's been work like the AI Bill of Rights, it's quite toothless even though I know that it was a big deal for the White House to release a draft blueprint for an AI Bill of Rights, but that has no money behind it, there's no enforcement behind it.
It's a document on the piece of paper.
They actually got Chips passed and like a lot of things, it passes under the rubric of competing with China and it's just at the company level, they're competing with each other and at the government level, they're competing with China.
To a point you've made, I think a couple of times, getting there quick is not obviously the right thing when there's a lot of danger from accelerating this even just beyond the point of human comprehension.
I just don't see that taken as an equity seriously enough there.
I think that the worries about China are a little bit fake in Silicon Valley, but I don't think they're fake at the national level, at the political level, and the power of that center of worry I think is actually very easy to underestimate.
Right now, the anti-China side of the political debate in Washington gets whatever it wants and what it wants is AI dominance, among other things.
That's potent.
Yeah, I think there's an unfortunate dynamic where you might expect the government and the national security apparatus to see these polls where machine learning researchers are like, yeah, 10% chance of the end of human civilization to be like, well, that sounds like a big threat to America's national security.
We should get on that.
If anybody's going to end the world, goddamn it, it's going to be America.
You treat threats as the kind of threat you've encountered before.
We've had a Cold War before, so they're ready to have a Cold War.
If China wants to have a Cold War, I think there's a lot of people that fits into the frameworks they have for understanding threats.
They have thought a lot about it.
They have a sense of what America winning in geopolitical competition with China looks like or takes or whatever, and they don't have any sense of what to do if, in fact, the

most destabilizing and dangerous technology is being developed here in America by patriotic Americans who want to either bring about cool sci-fi future or we can't guarantee a demon doesn't slip through the portal, but that's what we've got.
It feels, to me, very strange in some ways to be in these conversations where people are like, yeah, we are going to radically transform the entire world we live in pretty much overnight, and no one's very interested in making sure that that happens in a way that is safe for all of the existing people.
I think that where there is pressure coming from the national security establishment at all, like you said, it's coming in the direction of, yeah, beat China, and like I said, I don't like the CCP, but I don't like what the competition lens does to our odds of building AI that's good for anybody.
So we've talked here about a lot of things that could go wrong, though obviously there are many more.
I wonder before we end, talk a bit in a specific way about things that, particularly in the near term, could go right, not just in making these systems safer, but in things they could actually do that would be good, because right now it's still sometimes a little bit hard, I think, to visualize that, like, okay, having a chatty BT bot that can help you cheat on exams is cool, or I enjoy with my four-year-old, all sometimes called Dolly, and let him come up with pictures he would like the computer to draw.
I'm like, that's really fun.
But when you talk to these AI researchers, I mean, I think this is my worry about the bias of continuously quoting this 10% number about it might end the world.
They put a slightly higher, not as much higher, probabilities I would like, but a somewhat higher probability on it being very good for humanity than very bad.
Is the main thing you hear scientific advance?
I mean, do they have other visions here about what they're doing to make people's lives better?
I'm excited about translation.
There is so much of a barrier to being able to participate in intellectual life in so many fields.
If you don't speak English, there are so many brilliant people who, you know, work extremely hard as adults to learn English, to learn about fields they're interested in, to get jobs they're interested in, and I think we are on the brink of translation technology just changing the game there, where everybody who isn't in a country that's chosen to firewall itself has access in their native language, in whatever language they read most easily, to all of human knowledge.
I'm really excited for that.
The drug discovery stuff is legitimately cool and valuable.
It's very hard to predict protein folding, but it might be much easier to do with AI.
Similarly, I do think, I think it's super true what you're saying earlier, that science grounds out in wet labs that take time to do things, but I think there's still a lot of data processing and looking at images and stuff like that that is going to be incredibly valuable and save a lot of people's time to have AI do.
I think that while it will be extremely disruptive and while it would ideally happen slower,

automating a ton of human labor is fundamentally the sort of thing that could be really good.
It's mostly the busy work that we don't, in fact, like sending emails and writing applications and arranging meetings is nobody's dream job.
It could be really cool to have a much more productive economy in which a lot of that value is happening without people having to spend so much of their time on it.
I am fundamentally pretty in favor of the whole increased productivity and then stuff gets better.
There's some good stuff there with better code writers and better meeting schedulers and stuff like that.
Something I wonder there is also increased creativity.
I think there's been correctly a lot of concern and anger about these systems trained on the actual work of artists and writers and so on, and those people aren't compensated.
But I also think that when you, I can't draw anything, like I'm a terrible artist, my mom is actually was a professional artist.
I wish that I could draw things I can't.
It's neat for me that I can tell the computer what to draw and it will draw it.
It allows me to play around in art in a way I couldn't before.
Beyond that, it used to be that a lot of people could illustrate something, but very few people could create a video game.
Now a lot of people might have amazing ideas for video games and I think within a few years they're just going to be able to tell the computer how to make it with very detailed instructions.
Same thing with movies, same thing with music.
There is value to knowing how to do the underlying kind of material creation yourself.
I don't want to take anything away from that and I think that fosters forms of creativity that are not always possible if you don't understand the actual basics, but also I do think there's something to just opening up vast vistas of creation to people who we're never going to learn how to code, we're never going to learn how to use a synthesizer, but now maybe can work in tandem with a computer to create amazing art that they will just enjoy creating and other people might enjoy using.
I think that stuff is actually really cool.
Yeah, I think there is a lot of potential to really unlock amazing things and instead of replacing humans with aliens, just have really powerful assistive tools that can turn our imagination into reality more easily instead of writing an essay for you, being a really good editor and assistant that suggests a bunch of relevant stuff and helps you write a better one.
I think whether we do get that or not depends a lot on who has, as you said, financial incentive to build it, to make a really, really good tool that strengthens the people who wield it instead of replacing them.
But I think it's possible to build, and if it happens, I think it could be a really cool thing.
I think it's a good place to end, always a final question, which I know you've prepared for at length.
What are the three books you'd recommend to the audience?

Yeah, so I have been enjoying the making of the Atomic Bomb.
Obviously, the relevance to AI is very notable, but I think it's also just an interesting window into anything where a group of brilliant people are trying to make something go well and the combination of flaws and personal strengths, and I think a lot of tech companies from the outside, it's very hard to see why one succeeds and one fails, and it often comes down to these weird human factors, and it's terrifying to think that the development of one of the most powerful and transformative technologies in our world also had just so much of the weird human factors to it, but I think it really did, and it makes me think a lot about AI, but also just how stuff gets done.
The other thing I caught up on was Asterisk, which is a new magazine from some of the weird people that you like to talk about.
It's people in the AI and effective altruism and rationalist world.
I was reading about how we invented oral rehydration therapy, which is it's sugar and salt and water, took us until the 1960s, and the progress studies people always love the question like, why did it take us so long to get the bicycle?
Why did it take us so long to get the plow?
Like how much was somebody having a brilliant idea, and how much was like that there's actually a ton of execution details, and how much is like you actually need fairly advanced machining for bicycles.
It might not seem like it, but you do or whatever.
So oral rehydration therapy, right?
Why did it take us until the 1960s to figure out that 4 million kids a year were dying unnecessarily for a lack of sugar and salt and water?
I was obsessed for my entire teenagehood with the Silmarillion Tolkien's prequel to Lord of the Rings.
It's not even a good fiction book.
It reads something between a Bible and like a packet of world-building notes, but I was in love with it.
It's a beautiful insight into how someone puts a story together.
When the AIs can surpass Tolkien, that's when I'll know we're really in trouble because he was something else.
Kelsie Piper, thank you very much.
Thanks!
The Ezra Clanche was produced by Emma Fagagou and Galvin Jeff Gelb, Rochette Karma and Kristen Lin.
Fact-checking by Michelle Harris and Kate Sinclair, mixing by Jeff Gelb, original music by Isaac Jones, audience strategy by Shannon Busta.
The executive producer of New York Times' opinion audio is Andrew Ostrasser, and special thanks to Carol Sabarow and Christina Semilewski.