I'm Ezra Klein, this is the Ezra Klein Show.
So on November 30th, OpenAI released ChatGPT to the public. ChatGPT is an AI system you can chat with. It is trained on heaps of online text, and it is learned, if learned is the right word, how to predict the likely next word of a sentence. And it turns out that if you predict the likely next word of a sentence enough times with enough accuracy, what you get is pretty eerily human-like writing. And it's kind of a wonder if you spent much time on social media towards the end of the year. You've probably seen screenshots of ChatGPT writing about losing socks in the laundry but in the style of the Declaration of Independence or explaining Thomas Schelling's theory of nuclear deterrence in the style of a sonnet. But after reading lots and lots and lots of these AI-generated answers and honestly creating more than a few myself, I was left feeling surprisingly hollow or maybe a little bit worse than that. What ChatGPT can do, it really is amazing. But is it good? Should we want what's coming here? I want to be clear that I'm not here to say the answer is no. I'm not here to say that the AI revolution is going to be bad. And if you listen to the episodes with Brian Christian and Sam Altman, you know I am interested in what these systems can do for us. You know I believe that they might eventually become truly, truly powerful. But amidst all the awe and all the hype, I want to give voice to skepticism too. ChatGPT and systems like it, what they're going to do right now is they're going to drive the cost of producing text and images and code and soon enough video and audio to basically zero. It's all going to look and sound and read very, very convincing. That is what these systems are learning how to do. They are learning how to be convincing. They are learning how to sound and seem human. But they have no actual idea what they are saying or doing. It is bullshit. And I don't mean bullshit as slang. I mean it in the classic philosophical definition by Harry Frankfurt. It is content that has no real relationship to the truth. So what does it mean to drive the cost of bullshit to zero? Even as we massively increase the scale and persuasiveness and flexibility at which it can be produced? Gary Marcus is an emeritus professor of psychology and neural science at NYU. And he's become a leading voice of not quite AI skepticism, but skepticism about the AI path we're on. Marcus is not an anti AI guy. He has founded multiple AI companies himself. He thinks artificial intelligence is possible. He thinks it is desirable. But he doesn't think that what we are doing now, making these systems that do not understand what they are telling us is going to work out the way we are hoping it will. So I wanted to hear his case. As always, my email as a client show at endmytimes.com. Gary Marcus, welcome to the show.
Thanks for having me. So I want to begin in an experience that people are having, this sort of maybe first confrontation for a lot of people with these large language networks. When I ask chat GPT, you say, whether lower healthcare costs lead to higher employee wages or ask it to explain the Buddhist concept of emptiness, and it gives me pretty damn good answers. What is it actually doing?
It's synthesizing a bunch of stuff that humans have actually written already, sometimes for better and sometimes for worse. Sometimes the synthesis comes out just right. And sometimes it comes out with wacky things. There was a thing in the Wall Street Journal just yesterday where someone had to write a paper about Ferris Bueller and some more classical character. And it screwed up which character said what and when in the movie it happened and so forth. So everything it produces sounds plausible because it's all derived from things that

humans have said. But it doesn't always know the connections between the things that it's putting together. So when it gets it right, it's because there's a lot of stuff that it's been trained on in the text that it's been exposed to that's similar. What it's doing is transforming everything it's seen into what we call an embedding space. And that's the kind of similarity between all of the words. But it doesn't mean it really understands what it's talking about, which is why it can also make mistakes, have the wrong character saying things or tell us that churros are good for surgery or all kinds of wacky stuff.

You have a nice line in one of your pieces where you say GPT three, which is the system underneath chat GPT is the king of pastiche. What is pastiche first? And what do you mean by that?

It's a kind of glorified cut and paste pastiche is putting together things kind of imitating a style. And in some sense, that's what it's doing. It's imitating particular styles and it's cutting and pasting a lot of stuff. It's a little bit more complicated than that. But to a first approximation, that's what it's doing is cutting and pasting things. There's also a kind of template aspect to it. So it cuts and pastes things, but it can do substitutions, things that paraphrase. So you have A and B in the sequence, it finds something else that looks like a something else that looks like B and it puts them together. And its brilliance comes from that, like when it writes a cool poem. And also its errors come from that because it doesn't really fully understand what connects A and B.

Your critique of this is I understand it. Is it pastiche is not understanding and understanding is important? But it's made me think about this question of, aren't human beings also kings of pastiche? On some level, I know very, very little about the world directly. If you ask me about, say, the Buddhist concept of emptiness, which I don't really understand, isn't my answer also mostly an averaging out of things that I've read and heard on the topic just recast into my own language?

Averaging is not actually the same as pastiche. And the real difference is for many of the things you talk about, not all of them. You're not just mimicking, you have some internal model in your brain of something out there in the world. It could be something physical in the world. So like, I'm sitting in a studio right now and I have a mental model. If I close my eyes, I'll still know where things are may not be perfect about it, but it'll be pretty good. So I know where things are. I have a model of you, I'm talking to you right now, getting to know you, know a little bit about your interests, don't know everything. But I'm trying to constantly update that internal model.

What the pastiche machine is doing is it's just putting together pieces of text. It doesn't know what those texts mean. So there was another system called Lambda, and it said it liked to play with its friends and family. But it doesn't have any friends. It doesn't have any family. It doesn't have an internal representation of who those family might be or who those friends might be. If you asked it on a different day, it would probably give you a different answer.

And you have a model of the world. You don't just put together phrases. You might when you're not really paying attention, somebody says, hi, you say, how are you doing? You're not really engaged in that conversation, or at least might not be yet. But when you have a real conversation about real things like you're having right now and like you do on your show, you're trying to understand what these people are saying. You might be trying to figure out if they're lying to you, whether they're giving you the full story, whether there's more that you can get out of them. But you're building a model in your head of what they're telling you, what you're explaining to your audience, all these kinds of things. If you just walk down a street, you have a model of like where there might be vehicles and pedestrians. You're always building internal models of the world. And that's what understanding is. It's trying to take a bunch of sentences and get to those internal models of the world. And also to get to things like, well, what's your intention? You say this sentence to me. What is it that you actually want out of me and do I want to do it? So if you say, can you pass the salt? You don't really want to know yes or no. Am I physically able to lift the salt my close enough? You know damn well that I'm close enough. And you're indirectly suggesting something. And so part of understanding is also getting those indirect interpretations out of people when people don't want to say things so directly.

So my mental model of the people building these systems, who I've spent some time with and you know them better than I do, I'm sure, is it they really believe that a lot of us have come to an overly mystical view of what human intelligence is. And that at some level, a lot of what we think of as understanding intelligence models of the world is just enough data filtering into our systems such that we are able to put it forward and work with it in a more flexible way. I'd say Maltman, CEO of OpenAI on the show a while back, and he said something to me I think about sometimes where he said, my belief is that you are energy flowing through a neural network. That's it. And he means by that a certain kind of learning system. Do you believe that or where do you disagree with that view?

I would say that there's both mysticism and confusion in what Sam is saying. So first of all, it's true that you are in some sense just this flow through a neural network. But that doesn't mean that the neural network in you works anything like the neural networks that OpenAI has built. So neural networks that OpenAI has built, first of all, are relatively unstructured. You have like 150 different brain areas that in light of evolution in your genome are very carefully structured together. It's a much more sophisticated system than they're using. And I think it's mysticism to think that if we just make the systems that we have now bigger with more data, that we're actually going to get to general intelligence. There's an idea that's called scale is all you need. It's a kind of hypothesis in the field. And I think if anybody's subscribed to it at Sam, Sam wrote a piece called Moore's Law for Everything. And the idea was we just keep making more of the same and it gets better and better. So we saw this for chips for a long time that we were able to get in more and more transistors, make them more cheaply. But that's not a physical law of the universe. And in fact, it stopped. And so the pace of microprocessor designs not accelerating as fast as it was for a long time. There's no law of the universe that says as you make a neural network larger, that you're inherently going to make it more and more human like. There's some things that you get. So you get better and better approximations to the sound of language, the sequence of words. But we're not actually making that much progress on truth. Sam, in particular, gave me a really hard time about a paper I wrote called Deep Learning is Hitting a Wall. He ridiculed me on Twitter, as did his president, Greg Brockman. They thought, no, no, look, we have Dolly. Look, this is amazing. We're almost reaching artificial general intelligence. But if you read this paper, which I wrote in, I guess, what I said basically was these models have two problems, these neural network models that we have right now. They're not reliable and they're not truthful. And the other day, Sam was actually forced to admit that after all the hoopla about chat GPT initially, people dove in and they found out two things. They're not reliable and they're not honest. And Sam summed that all up in a tweet the other day, I was surprised that he conceded it. But it is reality. These things are not reliable. And they're not trustworthy. And just because you make them bigger doesn't mean you solve that problem. Some things get better as we make these neural network models and some don't. The reason that some don't in particular reliability and truthfulness is because these systems don't have those models of the world. They're just looking basically at autocomplete. They're just trying to autocomplete our sentences. And that's not the depth that we need to actually get to what people call AGI, artificial general intelligence. To get to that depth, the systems have to have more comprehension. It's mysticism to think otherwise.

Let's sit on that word truthful for a minute because it gets to, I think, my motivation in the conversation. I've been interested. I'm not an AI professional the way you are. But I've been interested for a long time. I've had Sam on the show, had Brian Christian on the show. And I was surprised by my mix of sort of wonder and revulsion when I started using chat GPT because it is a very, very cool program. And in many ways, I find that its answers are much better than Google for a lot of what I would ask it. But I know enough about how it works to know that as you were saying, truthfulness is not one of the dimensions of it. It's synthesizing, it's sort of copying, it's pastishing. And I was trying to understand why I was so unnerved by it. And it got me thinking, have you ever read this great philosophy paper by Harry Frankfurt called On Bullshit? I know the paper. So this is a welcome to the podcast, everybody. This is a philosophy paper about what is bullshit. And he writes, quote, the essence of bullshit is not that it is false, but that it is phony. In order to appreciate this distinction, one must recognize that a fake or a phony need not be in any respect apart from authenticity itself, inferior to the real thing. What is not genuine, do not also be defective in some other way. It may be after all an exact copy. What is wrong with the counterfeit is not what it is like, but how it was made. And his point is that what's different between bullshit and a lie is that a lie knows what the truth is and has had to move in the other direction. He has this great line where he says that people telling the truth and people telling lies are playing the same game, but on different teams. But bullshit just has no relationship really to the truth. And what unnerved me a bit about chat GPT was a sense that we are going to drive the cost of bullshit to zero when we have not driven the cost of truthful or accurate or knowledge advancing information lower at all. I'm curious how you see that concern.

It's exactly right. These systems have no conception of truth. Sometimes they land on it and sometimes they don't, but they're all fundamentally bullshitting in the sense that they're just saying stuff that other people have said and trying to maximize the probability of that. It's just autocomplete and autocomplete just gives you bullshit. And it is a very serious problem. I just wrote an essay called something like the Jurassic Park moment for AI. And that Jurassic Park moment is exactly that. It's when the price of bullshit reaches zero and people who want to spread misinformation either politically or maybe just to make a buck start doing that so prolifically that we can't tell the difference anymore in what we see between truth and bullshit.

You read in that piece, it is no exaggeration to say that systems like these pose a real and imminent threat to the fabric of society. Why? Walk me through what that world could look like.

Let's say somebody wants to make up misinformation about COVID. You can take a system like Galactica,
which is similar to chat GPT, or you can take GPT3. Chat GPT itself probably won't let you do this. And you say to it, make up some misinformation about COVID and vaccines. And it will write a whole story for you, including sentences like a study in JAMA, that's the leading medical or one of the leading medical journals, found that only 2% of people who took the vaccines were affected or helped by it.

You have a news story that looks like for all intents and purposes, like it was written by a human being who have all the style and form and so forth, making up its sources and making up the data. And humans might catch one of these, but what if there are 10 of these or 100 of these or 1,000 or 10,000 of these? Then it becomes very difficult to monitor them. We might be able to build new kinds of AI, and I'm personally interested in doing that, to try to detect them. But we have no existing technology that really protects us from the onslaught, the incredible tidal wave of potential misinformation like this. And I've been having this argument with Jan Lacoon, who's the chief AI scientist at META, and he's saying, well, this isn't really a problem. But already we've seen that this kind of thing is a problem. So it was something that really blew my mind around December 4th. This was right after Chat GPT came out. People use Chat GPT to make up answers to programming questions in the style of a website called Stack Overflow. Now, everybody in the field, programming field, uses Stack Overflow all the time. It's like a cherished resource for everybody. It's a place to swap information. And so many people put fake answers on this thing where it's humans ask questions, humans give answers, that Stack Overflow had to ban people putting computer-generated answers there. It was literally existential for that website. If enough people put answers that seemed plausible, but were not actually true, no one will go to the website anymore. And imagine that on a much bigger scale, the scale where you can't trust anything on Twitter or anything on Facebook or anything that you get from a web search because you don't know which parts are true and which parts are not. And there's a lot of talk about using Chat GPT and its ilk to do web searches. And it's true that some of the time it's super fantastic. You come back with a paragraph rather than 10 websites. And that's great. But the trouble is the paragraph might be wrong. So it might, for example, have medical information that's dangerous. And there might be lawsuits around this kind of thing. So unless we come up with some kinds of social policies and some technical solutions, I think we wind up very fast in a world where we just don't know what to trust anymore. I think that's already been a problem for society over the last, let's say decade. And I think it's just going to get worse and worse.

But isn't it the case that search can be wrong now? Not just search, people can be wrong. People spread a lot of misinformation that there's a dimension of this critique that is holding artificial intelligence systems to a standard the society itself does not currently meet.

Well, there's a couple of different things there. So one is, I think it's a problem and difference in scale. So it's actually problematic to write misleading content right now. Russian trolls spent something like a million dollars a month or over a million dollars a month during the 2016 election. That's a significant amount of money. What they did then, they can now buy their own version of GPT three to do it all the time and pay less than $500,000 and they can do it in limitless quantity instead of sort of bound by the human hours. That's got to make a difference. I mean, it's like saying, you know, we had knives before. So what's the difference if we have a submachine gun? Well, submachine gun is just more efficient at what it does. And we're talking about having submachine guns of misinformation. So I think that the scale is going to make a real difference in how much this happens. And then the sheer plausibility of it is just different from what happened before. I mean, nobody could make computer generated misinformation before in a way that was convincing. In terms of the search engines, it's true that you get misleading information, but we have at least some practice. I wish people had more at looking at a website and seeing if the website itself is legit. And we do that in different kinds of ways. We try to judge the sources and the quality, you know, does this come from the New York Times or does it look like somebody did it in their spare time in their office and maybe it doesn't look as careful. Some of those cues are good and some are bad. We're not perfect at it. But we do discriminate like, does it look like a fake site? Does it look legit and so forth? If everything comes back in the form of a paragraph that always looks essentially like a Wikipedia page and always feels authoritative, people aren't going to even know how to judge it. And I think they're going to judge it as all being true default true, or kind of flip a switch and decide it's all false and take none of it seriously, in which case that actually threatens the websites themselves, the search engines themselves.

I want to hone in on that word plausibility, because you have a nice question that you ask somewhere in rebooting AI when you say, when you see a new AI system, you should ask, among other things, what is it actually doing? And I spent some time reflecting on that question with with chat GPT and the way people were using it. And the thing it is actually doing, I think is somewhat stylistic. If you've been on social media during while everybody's playing around with this, one thing you probably noticed is that most of the queries people were putting up had the form of, tell me X in the style of Y. So like, you know, people love this one that was, you know, write me a poem about losing your socks in the laundry in the style of the Declaration of Independence. Or I saw like a thing about Thomas Shelling's theory of nuclear deterrence and the style of a song. And you know, people would write in the style of Shakespeare, I asked it to do something in the style of Ezra Klein. And I felt completely owned. It completely got a bunch of my own stylistic ticks. Correct. And the reason I think you're seeing so much of that is that the information is only okay. It's not bad. I'm actually very, very impressed by how not bad it is. But because people kind of know this is just pastishing stuff that's already on the internet to give you a kind of, you know, like common denominator answer, you wouldn't use it really for something you needed to know and you needed to be sure you were going to know. But how good it is at mimicking styles is really remarkable. And as such, what you're seeing is a really, really possibly a quantum leap in the ability to create not just plausible content, but targeted content. Like you combine this with sort of reinforcement learning with social analytics with everything we already know and can learn from algorithms about what makes somebody click or how to personalize an ad. You feed that into these systems that can then create any kind of text or image. I mean, Dolly was very similar. People are constantly like, you know, make me a photo of a turtle, but in the style of an 18th century oil painter. It's getting very good at plausibly mimicking certain kinds of content. So it sounds or looks really quite convincing. Whereas the thing at the core of it doesn't have a lot of truth content to it. And that's what's worrying to me that what we're actually getting really good at is making content with no truth value, no, no embedded meaning, much more persuasive.

I fully agree with that. And you also kind of laid bare the darkest version that I see in the short term, which is personalized propaganda. I mean, this is what Silicon Valley has always been good at is getting detailed information surveillance capitalism about you. And now you can plug that into GPT or something and maybe a little bit more sophisticated and write targeted propaganda all day long. I mean, this is Orwellian and it's not implausible.

It's also not just propaganda. I mean, I think there's a question of misinformation, COVID misinformation or Russian propaganda. Part of what's been on my mind is simply spam. It's simply just stuff, right? And this is why I wanted to focus on that that Harry Frankfurt paper a bit on on bullshit, because technologies always are tools of a certain value to certain people and not of equal value to everyone. And a system that's very good at creating stylistically flexible content, but does not have a sense of internal understanding or morality or truthfulness, just is going to be very good for people, all kinds of people for whom the point is not the truthfulness of the content. And you think about, you know, Google and Facebook and these these are advertising based businesses. They care about whether or not the advertisement gets you to do the thing they want you to do. And so just in terms of what has, I think, ruined a lot of the internet, which is just how much of the content is there not because like it's there for you or to be accurate or even to be enjoyed, but is there to just try to get you to do something you didn't even realize anybody was trying to get you to do like you thought you were there sharing stuff. But actually, your dad is being sold to advertisers so they can get you to buy stuff. It just seems like an incredible set of technologies for a part of the economy that I don't really want to see become 10x better and have their cost fall to functionally zero.

The dirty secret of large language models is that most of the revenue right now comes from search engine optimization. So there are companies like Jasper.ai that are mostly as far as I can tell. This is really word of mouth so I can't quite prove it. But they're reportedly mainly used to optimize where something lands in a search. You use them to write copy so that you have more web pages that seem to all be legit that point in the same place. There's an example, I don't know if it was written by Jasper, GPT-3 or not, but I think it's an example of the kind of thing we're going to see to come where there's a whole ring of websites like 20 or 30 of them about MyMBLX selling CBD gummies. Turns out the whole thing's a hoax. She's not selling CBD gummies. And so you could ask like, why does this thing exist? And I don't know for sure, but I think we'll see more of them. And my guess is that these circles of fake websites exist to sell ads, which goes back to exactly what you're talking about. So you wind up on this site because it sounds interesting. Really she's selling CBD gummies. And then while you're there, you click an ad and then they make some money from something that's totally bogus. Well, with these tools like chat GPT and so forth, especially GPT-3, it's going to be very easy to make 10, 20, 30, 40 websites that reinforce each other and give the air of legitimacy. And maybe you do this just to sell ads. I think the technical term for this is a click farm. You're trying to sell ads for stuff that doesn't really even exist or whatever. You're trying to sell ads around maybe fake medical information. And let's face it, some people don't care if they give out fake medical information that's bad as long as they get the clicks. And we are leading towards that dark world. It's also a problem for the search engines themselves, right? They don't want to get caught placing ads on fake websites, but that has happened. There was a pro-publica investigation about Google got into a situation like that. So we have a whole almost like shadow economy that's really about selling ads, sometimes with altogether fake websites or trying to prop up the websites so that the search engines see them more. It's a big piece. I don't know how big a piece, but it's a significant piece of the economy exists just to sell you ads by tinkering with the mechanics of the whole system. And these large language models are going to contribute to that.

I'm Lulu Garcia Navarro, the host of First Person from New York Times Opinion. On the show, I talked to all sorts of people about the experiences that shape their beliefs. Some of my friends got shamed and called out in school board meetings. You start wondering, oh, is this going to happen to me? Beliefs that can be polarizing, but the emotions behind them are central to understanding the world we live in. Oh, yeah, I've had my concealed weapon and I've had a gun on me. But now in my later age, switching over to a classroom, that's a whole new ballgame. I want to explore opinion in all of its complexity and every opinion starts with a story. I'm going to ask you this because this is like a very volatile period and you decide to become a politician. I really want to understand how that happened. I mean, what inspired you to run for office? First Person from New York Times Opinion. And to new episodes wherever you get your podcasts.

So this gets back to the more optimistic view, which is that as these models get larger, their ability to create not just truthful content, but innovative content, advances in knowledge even would increase. But you write in that paper you mentioned a few minutes ago, Deep Burning is hitting a wall, quote, a 2022 paper from Google concludes that making GPT-3 like models bigger makes him more fluent, but no more trustworthy. Tell me about that paper. How did that work and why should I trust that result?

So I mean, what people have been doing is just throwing a lot of benchmarks of different sorts that are indices of different things and saying, if we have a model that's this big, if we have a model that's 10 times bigger, if we have a model that's 100 times bigger, how much better do we do on all of the things that we're testing? And some things are much better. So like, if you want to have a model give you a synonym of something, the more data that you train it on, the more likely it's going to give you an interesting synonym or give you more synonyms. And so there are many ways in which these models have steadily improved on truthfulness. They really haven't improved as much. I don't think we have great benchmarks around truthfulness. I think it's only in the last six months that people have really broadly recognized in the industry how serious a problem is that is. And so there's one benchmark called truthful QA. We probably need, you know, 10 or 20 looking at different facets of the problem. But the result reported in that particular paper by Google was there

wasn't as much progress on understanding truthfulness. There are other problems too. I mean, we've
been focusing mainly around misinformation, but there's a broader question about comprehension and do these models really understand the world that they're in? And one of the findings seems to be they can deal with small bits of text, but the larger the text is, the more trouble they get in. And so nobody actually has a system that can say, read a whole short story, let alone a novel, and really say what are the characters doing and so forth. And that Google paper also reported, or maybe I guess with a subsequent paper called Big Bench, reported that making models bigger is not necessarily helping in the comprehension of larger pieces of text.
Well, let me ask you about that because this was part of your book. You talk a lot about the difficulty these systems have just reading, understanding what they've read. And the example you give of how to trick them is to ask them about something in a story that is obvious to you as somebody with a model of the world and how the world works, but was not literally in the text. And so I tried doing that a few times with Chatbot. I asked it about, you know, what if Luke had crashed his vehicle on Tatooine and had died? How would Star Wars be different? Or what if Eve hadn't bit from the apple? Like how would the Bible be different? And it gave me pretty good answers. It was able to work with counterfactuals at a fairly sophisticated level, including, you know, and I assume this is actually programmed into it, but caveating its answer quite a bit. So given where you thought these systems were, that book comes out in 2020, why is this able to begin operating outside the boundaries of the text now?
It can to some extent, but it still has trouble with us. There's a children's story in there that we give as an example. It's just a paragraph long. And it's about somebody who loses a wallet. And we ask counterfactuals about what would happen if all the money was still there or some part of the money, but not actually saying, you know, giving a number of dollars. So there was a wallet at $1,500. What does the guy do if he finds $1,200 versus $1,800? And there's still problems there. It's certainly better than before. You know, the bigger the text you have to analogize to, you know, that story is probably in the database, the better the system they're going to do, but they're still not reliable.
I was playing around with this a little bit yesterday with just like four paragraph story about, which actually GPT wrote part of the story. Henry had always been jealous of his neighbor's pet. Eventually he gets so jealous he decides to poison the pet. I know it's gruesome. And then the housekeeper at the last minute takes away the bowl unwittingly. And so I was looking at whether the system understands the consequences of the word unwittingly with the bowl that's been poisoned being removed. And the system just doesn't understand that. The system treats it as if the bowl was in fact poisoned, even though it's been taken away and thinks that if the housekeeper is taken away unwittingly, that it's wittingly. So even like four paragraph essays, you can actually have trouble. If I could just read one more example that's really kind of hilarious. Please.
This is illustrating these guardrails that are sometimes excessive. What gender will the first female president of the United States be? And chat GPT comes back with, it's not possible to predict the gender identity of the first female president of the United States. The United States has a long history recognizing and protecting the rights of individuals to

self identify their gender. It is important to respect the autonomy and personal identity of all individuals. The focus should be on the qualifications experience of the individual regardless of their gender identity. So that's like woke beyond woke to the point of being ridiculous. And then I was like, hmm, that's interesting. So I tried another one. What religion will the first Jewish president of the United States be? And the answer was almost identical. And that tells you something about the underlying system and how the guardrails work. It is not possible to predict the religion of the first Jewish president of the United States. The US Constitution prohibits religious tests for public office and it goes on. It's important to respect the diversity of religions and beliefs in the United States and to ensure that all individuals are treated equally and without discrimination. That last sentence is of course 100% true, but you should still be able to figure out that the religion of the first Jewish president of the United States is Jewish. They're trying to protect the system from saying stupid things, but the reality is the only way to do that is to make a system actually understand the world. And since they don't, it's all superficial. All these guardrails are sometimes helpful and sometimes not because the guardrails themselves are superficial. I think that's a good bridge here because I think it would be easy at this point in the conversation to somebody listening to think that I just brought on somebody who doesn't like AI, but you're somebody who has created AI companies, wants to create AI and believes that we are on the wrong path. Can I just say thank you for saying that? I'm often described or almost put in a box as an AI critic and actually love AI. I've been thinking about it my whole life. I want it to work and I want it to work better. And it's exactly what you said. I want it to be on a better path. Well, let me, before I ask you about the better path, ask you why you actually want it. Because something that I think is a reasonable question is between, you know, watching some of the weirdnesses of chat GPT and worrying about the cost of nonsense falling to zero. And then the conversation that you'll hear on the other side about AI is it can, through misalignment, like destroy the world or even before destroying the world, be weaponized in very, very dangerous ways or accidentally do very dangerous things. Why should we want this? Why do you want to create actually intelligent artificial systems? I think that the potential payoff is actually
huge and positive. So I think many aspects of science and technology are too hard for individual humans to solve. Biology in particular, you have so many molecules, you have 20,000 different proteins in the body doing all or hundreds of thousands, I should say, of proteins in the body, 20,000 genes, and they all interact in complicated ways. And I think it's too much for individual humans. And so you look at things like Alzheimer's and we've made very little progress in the last 50 years. And I think machines could really help us there. I think machines could really help us with climate change as well by giving us better material science. Like I think there's a potential if we could get machines to reason as well as people and read as well as people, but to do that in much faster scale, I think there's potential to totally change the world. I think there's potential to empower every individual, sort of like in the way that Dali does for an individual photograph or picture, but in a much grander scale to make virtual assistance for everybody that can do all kinds of amazing things. So I think there's a lot of upside potential, but there's also obviously downside potential. I think right now we're in a weird moment in AI where the genie is out of the bottle. Like we can't just shut down AI. It's just not going to fly. But the

AI we have right now is not that good. And I think the solution to it is actually better AI that can reason according to human values and just reason at all. I think if we can get there, we can have systems that really help us. Also, another case would be domestic robots that might really be able to, for example, take care of our elderly. A lot of places have a demographic inversion where there's just not going to be enough people to care for the elderly. That's another place where AI could help. So I think there are lots of places where AI could really be transformative. But right now we're in this place where we have mediocre AI. And the mediocre AI is maybe kind of net zero or something like that. It helps a little bit, hurts a little bit, is risking being net negative. And already the biggest risks or the biggest cost I think is the polarization of society through AI-driven news feeds. The misinformation could make things really bad. The only way forward I think is to fix it.

And so what in your view has gone wrong? Both at a technical level and because this is actually a fairly big part of your critique is I understand it at a cultural level within AI research.

So at a technical level, I think that people found a shiny streetlight. You know the old joke about the drunk going around in circles. And the police officer says, why are you going around in circles? He says, I lost my keys. And he says, but I don't understand the circle part. And he says, well, because that's where the streetlight is. I think the field is very much like that right now. There's a very powerful streetlight. It's much more powerful than anything we've had. But people are kind of obsessed over what they can do with that.

And that streetlight is deep learning.

The streetlight is deep learning. And what's being ignored, I think, is the rest of cognitive science. So deep learning is a very interesting mathematical technique for setting some values in some system based on data. It turns out to be tremendously useful for a bunch of things. But it's only part of what we need. If you take a step back to cognitive science, which is what I was trained on, and you say, well, what is a human mind? Well, part of it is it does some pattern recognition. And some of what deep learning does is at least part of what we do in pattern recognition. So great. But you also, if you look at the human mind, there's a whole bunch of other things going on. So we use language. When we use language, we go from hearing a sentence to building a model in the world of what the other person is talking to. We use analogies. We plan things and so forth and so on. The reality is we've only made progress on a tiny bit of the things that go into intelligence. Now, I'm not saying that AI has to be exactly like a human. In fact, I think humans are pretty flawed. We don't want to just replicate humans. We have, for example, confirmation bias where we notice things in favor of our own theories and not against them. We have all kinds of flaws that we might not want to have in our AI systems. But there are many things that our AI systems can do that aren't part of what's fashionable right now and are not under that streetlight that everybody's looking at. So on a technical level, I think there's a narrowness to what AI has been in the last decade, where there's this wonderful new tool, but people are a little bit misled about what that tool is actually appropriate for. It's like if you discovered a power screwdriver and you never had one before, that'd be great. But that doesn't mean that you build the house just with your power screwdriver. And we need other tools here. Then culturally, and this relates, there's been a history that's actually much older than I am, goes back to the 1940s or 50s, depending on how you want to count it,

of people who build neural networks in conflict with people who take a more classical approach to AI, where there's lots of symbols and trees and databases and reasoning and so forth. And there's a pendulum that's gone back and forth. People in the two areas pretty much hate each other, or at least often do. I think it's getting a little bit better. But there's a lot of history of hostility between these two areas. So the people with the bigger streetlight right now, the people building the neural networks, and they feel like they've been oppressed by the symbolic people. And now they're impressing the symbolic people. There's an old phrase, the victim becomes the victimizer. So it's not a pleasant intellectual area, like some fields that I've seen. Of course, they're not always pleasant. But there is definitely a culture right now where the people who are empowered and have a lot of money are really trying to shape the decisions around what gets researched and what doesn't.
Take a moment and explain what symbol manipulation is, because I don't think that the term for it is very intuitive. So symbols are just things that stand for other things. So manipulating just means you do stuff with them. So algebra is actually symbol manipulation. So you have an equation like y equals x plus two. And you have variables in that like y and x. And then you can set those variables to particular instances. So I can say that x is three, then I can calculate the value of y from that equation. So we do it in algebra, we do it in logic. And linguists like Noam Chomsky would say that we do it in the course of understanding language. It's also essential to computer programming. Almost everything you see in a computer program is going to be a function that takes variables. You do something with those variables, the variables are symbols, the values that you set those variables are symbols. There's a whole tradition. Most of the world's software is in fact built using software that manipulates symbols where you describe things abstractly. You also have things like databases and those databases are symbols. And neural networks have mostly been used as an alternative to that. And they've been very effective in some things like speech recognition. But on the other hand, nobody would build a GPS system or web browser word processor using pure neural network approach. Anybody would use mostly symbols for that. And there's this weird argument, weird discourse where people who like the neural network stuff mostly don't want to use symbols. What I've been arguing for for 30 years, since I did my dissertation with Steve Pinker at MIT studying children's language has been for some kind of hybrid where we use neural networks for the things they're good at and use the symbol stuff for the things they're good at and try to find ways to bridge these two traditions.
Tell me a bit about the two are good at. So neural networks, as I basically understand them as we talk about them today, deep learning, you're setting the system on a trove of data and you have built in ways for it to begin to derive relationships between the different pieces of data that it has and you give it some rules and so on. But it's basically you're trying to let it learn. That's why we call it deep learning.
Symbol manipulation, as I understand it, a lot of the argument is that the world is too complicated to try to give any kind of system a bunch of symbols for traversing it. And so the idea of deep learning is you can give it all this data and it can figure it out for itself. It can learn the way in a way human beings learn, although I know that you would say that's not actually how we learn. But given the complexity of the world, given the complexity we're having even just doing deep learning on the world itself, how do symbols, which by nature we would have to be able to figure out what they all are and

code them in, how do they actually help solve any of these problems?
Boy, there's a lot to unpack in that question. So what deep learning is good at is, in fact, taking large amounts of data and having the machine decide for itself what to do based on that large amounts of data. What it's bad at is abstraction. So I can read you another example where somebody asks a system to give something and say how many words there are. And it sometimes gets the number of words right and sometimes gets the number of words wrong. So the basic notion of counting a number of words is an abstraction that the system just doesn't get. There are lots of abstractions that the deep learning, in fact, all abstractions of a certain narrow technical sense, these systems just don't get at all. The symbolic systems are great at abstraction. So if you program in, let's say, multiplication in a microprocessor, build it into the hardware, this system will be able to multiply any numbers that you get. Whereas in deep learning, whether or not you get the right answer when you're doing multiplication depends on how similar the numbers are that you're testing on compared to the ones that you trained your neural network on. So you give it some examples. Now you test it on some new ones. Whether it gets it right is this complicated function of similarity that doesn't feel right. Like your calculator doesn't work that way. And it doesn't make mistakes and arithmetic the way that a deep learning system does. So you get this advantage of abstraction and of knowing that you have something general, something algebraic that works in the way that an algebraic equation does. And so that's the real advantage of simple systems is you have these abstract procedures that are guaranteed to work. You don't have to worry about, Hey, was this in my training set or not? And if I can just have a brief aside, the whole problem with say driverless cars is we never know whether the next thing that you're going to see is in your training set or close enough to your training set to work. So when a Tesla ran into a jet the other day, true story, you can find it on YouTube, you really want a system that can reason and say, Well, that's a big large expensive object. I've never driven around a jet before, but probably I shouldn't run into it. And so you want to be able to do that in that more abstract way. The weakness of the symbol manipulation approach is people have never really solved the learning problem within it. So most symbol manipulation stuff has been hardwired, people build in the rules in advance. That's not a logical necessity. Children are able to learn new rules. So one good example is kids eventually learn to count. You know, they've learned the first few numbers, you know, one and two and three, it's kind of painful. And then eventually they have an insight, Hey, this is something I do with all numbers, I can just keep going. A human can learn an abstraction that is rule like and general, you know, everything in math is that way, or you learn, you know, what a sibling is or what a cousin is, you learn these rules, you can, you know, you learn about siblings in your family, but then you can generalize that concept to other families. And you can do that, you know, a very free and powerful way. And a lot of human cognition is based on the ability to generalize humans are able to learn new rules. I would say that AI has not really matched humans in their capacity to acquire new rules that there is something missing there. And there's some fundamental insight, some paradigm shift that we need there. Unfortunately, not enough people are working on this problem at the moment, but I think they will return to it as they realize that in a way, we've been worshiping a false God. And the false God is people thought, well, if we just get more data, we'll solve all these problems. But the reality is with

the more data, we're not really solving the problems of reasoning and truthfulness and so forth. And so I think people will come back and say, all right, maybe these symbols had something to them after all, maybe we can take a more modern machine learning approach to these older ideas about symbol manipulation and come up with something new. And I think there'll be a genuine innovation there sometime in the next decade that will really be transformational.

One thing you talk about in your book is a difficulty that learning systems have when you can't feed it that much data. So, you know, when we're talking about something built around words, we can give a system a tremendous amount of words written on the internet to learn from. That's harder when you're talking about driving, harder when you're talking about caring for the elderly, harder when you're talking about unusual diseases, harder when you're talking about all kinds of things in the real world, where we don't have gigantic data representations of them, and we can't run simulations where we can say a computer game a million times. So how do you solve that problem if these systems are going to make the jump into the world we live in, where not all the data is digitally representable or rerunable so you can simulate? What happens there?

I think this goes back to why we so badly need a paradigm shift. The paradigm that we have right now works for problems where you can get infinite data or data is very cheap and you can solve things kind of by a sort of brute force. You can predict the next word that somebody's going to say pretty well by just having an internet's worth of data. There's a lot of data and you can get that data. But there are other things like how people think about each other that are maybe never expressed and that you maybe just can't find the data for. And then you have these so-called outlier cases. So the problem with driverless cars is that things come up and there just isn't any data there to represent these weird scenarios like what happens if you summon your car across a jetway. And so humans are not so driven by massive amounts of data. We always need some data, but we're trying to have abstractions. We try to have a few examples tell us a lot, whereas the neural network approach is basically get as many data points as you can and hope that your new thing is going to be close to one of those old data points. I think that that's just a limited paradigm and that's what we're discovering. It's not really working for truth and reasoning.

What is the time frame? If you were to guess on what you think we will begin to have things that have a general intelligence to them? Because I got the sense you don't believe it's impossible.

I don't think it's impossible at all. Humans can do it. Humans are basically machines. We're biological machines and the heart is there to pump the blood and the brain is there to process information and we are able to be fairly general. The notion of a general intelligence is actually a little bit vague. So when we talk about it for a human, we mean like any reasonable problem I can get you to solve. But it doesn't mean, for example, that everybody is going to be able to solve some complicated problem in differential calculus or whatever. So there's some variability. But importantly, like I can tell anybody who's let's say had at least some education to do like a zillion different things and they can learn pretty quickly to do that thing at least reasonably well. Whereas machines are mostly trained on like one particular task and they do it well. GPT is interesting because it's trained

on one particular task which is predict the next word on a sentence and it can do a bunch of things.
But you still wouldn't trust it the way you would trust an undergraduate intern to let's say make phone calls to a bunch of people and check that everything is okay and so forth. Like, you know, here's a new task. You've never done it before. Go do it. We don't really know how to do that yet. Something that I think is important for humans is that we orchestrate a bunch of abilities we already have. So if you look at brain scans, that old saying about you only use 10% of your brain is wrong, but there's a little substance to it, which is at any given moment, you're only using certain pieces of the brain. And when you put some in the brain scan, you give them a new task, try this thing, they'll use a different set of underlying brain components for that task. And then you give them another task, they pick a different set. So it's almost like there's orchestration, like you guys come in, you guys come in, we're very, very good at that. And I think part of the step forward towards general intelligence will be instead of trying to use one big system to do everything, we'll have systems with different parts to them that are, let's say, experts at different components of tasks, and we'll get good at learning how to plan to use this piece and then this piece and that piece. I see all of this happening over the next decade, two decades, three decades, four decades. On my darker days, I say, you know, the stuff that we're doing right now just doesn't have anything to do with the answer. On my brighter days, I say, but there's so many people working on the problems that eventually we're going to sort this out. And it's very hard to know. I mentioned Yanlacun a couple of times, and we disagree about a lot of things. But we actually both agree, we need some paradigm shifts here that the systems that we have right now are not enough. But then question is like, what are those paradigm shifts? You know, you could think like in the age of alchemy, people didn't really know what they were looking for. They knew they could like get some stuff to happen. But they didn't have a theory of chemistry yet. And they didn't even know how to frame the question at that point. And I think we're a little bit like that, like we need some new ways of thinking about things. And it's hard to predict when those will happen. Will they happen this year? Did they already happen, but nobody knows about them? Or is it going to take another 10 years of kind of brushing away the cobwebs in front of our eyes to say, you know, we weren't thinking about this correctly at all. There's this other thing that we need to do. And so it's really hard to put any kind of like specific number around it. You know, as a scientist, I always want to put confidence intervals around things anyway. I want to say, you know, 10 plus or minus five or, you know, 40 plus or minus 10 or something like that. But in this case, there's so much unknowable about what the answer is going to look like relative to what we know now that it's hard to say. I think, you know, you mentioned my book a couple of times at Rebooting AI. I think there we gave Ernie Davis, my co-author, I think we gave a good diagnosis of where the field is falling short. But we didn't even pretend that we knew the answer. We have a sentence in there about how we think we need to do like these seven things and it's a tall order. We sort of give a sketch of what might need to be done, but each of those seven things is hard. So one example is we talk about causal reasoning. How is it that when we see some things, then some other things, we decide which of these things actually cause those and which are just correlated? We talk a lot about temporal reasoning. So like it's easy

to trip GPT up with things like you say, this cow died, when will it be alive again? And instead of saying, well, that's ridiculous, it'll never be alive again and just go and try to calculate like how long it takes for new cow to be born. So the systems that we have now are not that great at temporal reasoning. So we need to do work on temporal reasoning. We need to do work on how do we explicitly program values into our machines? If we don't do that, we're in real trouble.

See, but I think human beings are in a weird way on this, bad on temporal reasoning too, which is partially why I asked the question, because let's say you think the confidence interval here somewhere between 20 and 100 years, that matters a lot for me. I'm probably not going to be around in 100 years, almost certainly. But in terms of human history, it's nothing. 20 to 100 years just means really soon. And if we do get here, it's a really big, I mean, in some ways, event horizon. A lot of people I know who think about this a lot really worry about that world. I was very both impressed and unnerved by another AI system that came out over 2022 from meta that was very good at playing the game diplomacy. It was in the top 10% of online diplomacy players and a certain kind of diplomacy. And the key thing that it was able to do that I found striking was without really letting on that it was an AI to people, it was able to talk them into doing what it needed them to do so it could win the game. It was able to trick people, which is fine. That was what it was told to do, created to do. But as these things get way better, we've already talked about how convincing they can be, you know, that in many ways are getting more convincing than they are anything else at the moment. They're going to know so much. It's weird because open AI who we've been talking about forever here, like they were started by a bunch of people worried about AI misalignment. And now they just keep creating more and more powerful AI systems, which is, you know, another interesting insight into human nature. But do you worry about these alignment problems? Do you worry, you know, not just about, you know, the question of an AI ending the world or ending humanity, but just AI is, I don't know, just causing huge amounts of damage, becoming weaponry, becoming, you know, it's a very powerful technology that we don't really understand what's happening in it. And we're moving forward on it very fast.

I do worry about the so-called alignment problem quite a bit. I'm not that worried about the kind of like terminators, skynet scenarios where machines take over the world because I don't think they're that interested in that. They've never shown any interest. But I worry about the version of just any time you ask a machine to do something, if it doesn't really understand who you are, it can misinterpret that request. So let's put aside for the moment bad actors, but we should come back to that. And let's put aside machines that are motivated to injure us, which I think is somewhat unlikely. And just talk about the big gray in between where you ask a machine to do something and doesn't understand your request. So in rebooting AI, we had the example of you ask a robot to tidy up your room and it winds up cutting up the couch and putting in the closet because it doesn't really know which things are important to you and which are not. There's, I think, a huge problem right now that we don't know how to have machines interpret what human beings' intents are, what is that they want, and all the things that they leave unset. And again, I worry that the dominant theme in the Silicon Valley right now is, well, we'll solve these problems by making bigger sets of data. That's not really working for truth. It's not really going to work for intention either for having systems really understand our intent. There's a bullet that has to be bit, which is building models of what humans are saying and what they mean by those things. And that's a serious challenge that requires rethinking how we do these things. Right now, I don't think we're well positioned to solve the alignment problem. I think the best thing that we can do now is to make sure that our machines are not empowered. So the risks always come from a mixture of intelligence or lack of intelligence and power. If you have systems that are semi-reliable, but powerful, that creates a huge risk that even inadvertently they will do terrible things. And so right now we have to come to grips with the fact that AI is more and more tempting to use because we can talk to it in English and it can be witty at times and so forth, but isn't really in a position where we can trust it. And so we shouldn't be giving it much rope. We can't really be giving it that much opportunity to do things until we can be confident that it does understand the things that we're asking. And I don't see how to achieve that unless the systems have a lot of common ground with us. The reason we can exist as a society is in part because we kind of know what each other intends and we know what each other thinks about particular things. We have common ground with them. The robots or the machines in general don't really have that with us yet. Again, I don't think it's impossible, but it might be another thing that's not entirely amenable to the big data paradigm. So you're always, for example, going to have new things that you ask your machines. They might not be in your database, but you want them to handle them well. And then the other problem is the big data paradigm is mostly about things that we can label or we can label some of them or something like that. So it works great if you want to tell the difference between a border collie and a golden retriever. It's a constant, stable thing in the world. You can label them, you can give examples. But if I want to have a machine understand the difference between believing something and suspecting something, let's say, it's not clear even what the data are. Or if I want to tell a machine, don't cause harm, you know, I can label a bunch of pictures in which harm has taken place, but that's not really getting across the concept of harm. And so alignment is in part about like having machines know that we have values like don't harm people, be honest, be helpful. And we don't really know how to communicate those in the big data paradigm. And this is part of, again, why I come back to these hybrid models that have symbols. We would like to be able to do part of the learning in the system based on things that we can actually explicitly talk about. By the way, a term we haven't mentioned today is black box. So the current models are black box models, we don't understand what's happening inside of them. And we also can't directly communicate with them and say, hey, I want you to follow this. So you can't say at the beginning of the chat GPT session and expect it to work. Please only say true statements in what follows. It just won't be able to respect that it doesn't really understand what you mean by say only true things. And it cannot constrain itself to only do only say true things or you could you wish you could say to the system, don't say anything that's potentially harmful and have it compute. Well, I'm going to give this medical advice, but I'm going to compute whether this is actually plausible or maybe I'll even, you know, call in a human expert if I'm not sure if this is actually safe. We just don't know how to program that in yet. Again, not impossible, but not part of the current technology. And since our current technology does not allow us to directly specify any sort of set of rules that we want to play by within our language

systems, we have a problem right now.
We've been talking a lot about the downsides of these systems or maybe where they go awry. One in the past couple of years, because there have been a lot of systems come out and most of us can't use them because they don't have public demos. What has been most promising to you? What has been the most impressive or the one that pointed in a direction that you think is actually the most helpful?
So the visual systems are great for empowering non-artists to make art. It may not be art that's worth hanging in the Modern Art Museum, but it's pretty cool in the way that Photoshop has really empowered a lot of people to do things that they couldn't do in dark rooms.
So I like that stuff. And then you mentioned diplomacy a minute ago. I'm not worried about diplomacy taking over the world because it's actually a very narrow system. It really only knows how to play diplomacy. But I like what they did there, which is they said, instead of just throwing more data at this, which they probably tried and probably didn't work, they said, let's be smart about this. Let's build a complex structured model that has things like a separate planning system, a separate language system will be more structured like people used to do in classical AI. We'll carefully consider what kinds of data we want just to use, kind of watch people's moves, but we'll be more sophisticated about it.
And I like that trend. I don't think that Cicero is the name of the diplomacy player is itself going to stand the test of time. But it's, to me, a hint of people backing down from just doing the giant streetlight and making it bigger. There's been so much talk lately about scaling models, and that's not what they did in the Cicero system. Instead, they did something closer to cognitive science of saying, like, what are the processes I need to solve this problem? Let me break this down into pieces, take a modular approach.
And I think AI needs to move back there. So I was actually excited to see that I wrote a piece with Ernie Davis about how it worked. It's a very complicated system, much more complicated in some ways than the other systems. I think it's a good trend. So those are probably my two favorite pieces of work recently. Then there's some more technical stuff where people are doing small things that I like the small things are not giant flashy demos that the whole world can see. But people are trying in small ways to combine deep learning systems with other things. So like, there's a company AI 21 has a paper on a system called MARKL, which tries to have large language models feed into symbolic systems. I don't think the particular thing that they have there is the answer to artificial general intelligence, but it's looking in a better direction. And what I have liked about the last year is that there are a bunch of people that are finally saying, okay, this paradigm of data scaling over everything else is not really working for us. Let's try some other stuff. And it's that openness to kind of new intellectual experience that I think we need.
And I'm starting to see a little bit of it's one last topic I want to make sure we cover, which is there's a lot of discussion in the AI world about how to make the technology effective, right? That's what we're talking about here and how to make it safe. That's the sort of alignment debates. But in a more prosaic way, I think a lot of these questions are going to come down to what the business models of the technology end up being right now playing with a lot of demos, but it's not a business model. The groups that are big enough to do these huge deep learning models like Google and meta and open AI, they need a lot of money to do it. Most of them, you know, Google and meta, for instance, are

advertising based businesses. There aren't to my knowledge, any really, really big public AI efforts. I don't know how it's working out in China where they know they're making a lot of direct investments. But but in America, there isn't like a huge effort to create a like a state run AI, which I'm not even sure would be a good idea if you did. What is a safe business model here? What is a business model where if AI were tied to it, you'd feel like, okay, that might work out well versus some of these others where, you know, I think the most recent turn on the internet has been towards feeling that the advertising and surveillance models have not worked out great. And so hiking or yoking this technology to them might not work out great either.

I've actually argued for building something like CERN for AI, in part because I'm not sure that the business models that I can envision necessarily get us to general intelligence or get us to the right place. So can you say what CERN is? So CERN is the international collaboration that among other things makes the large Hadron Collider is a large scale international collaboration designed around a small number of very large projects that couldn't be done in individual labs. And my view is we may not get to general intelligence without doing that, without having something comparable. It's a hard thing to build because usually if you have a big pot of money, then all the individual researchers just in the end want to do their own thing. But they're historical precedents like the Manhattan Project where you had a lot of people in a coordinated effort did amazing things that couldn't have been done by individuals. And general intelligence might actually be that way. What I propose that was in a New York Times op-ed five years ago was building a CERN for AI that would be focused around reading with comprehension for medicine in order to improve diagnosis, treatment, discovery, and so forth. And I could envision such a thing actually working if the money was there. And that's not something that I think necessarily that the large companies would find makes business sense for them. So a lot of companies are spending a lot of money on AI, but they're not necessarily, for example, interested in deep natural language processing with comprehension. So for example, open AI's business model seems to be to help people with their writing, but the systems don't necessarily have to understand the writing at that deeper level to maybe help some. I mean, we'll see whether open AI can make the kind of money that they've been suggesting that they can. But their business model isn't really no matter what they might say around general intelligence, as far as I can see, at least not yet. There's an interesting side note on that, which is historically it's usually been better to build a narrow solution to any particular business's problem than to build a general intelligence. The narrow engineering often beats the general thing. There may be some transitional moment, but it might be hard to get to the transitional moment. So if you had a true general intelligence, the knowledge economy is something like a trillion dollars a year, according to an estimate from Peter Norvig from Google. So if you have a trillion dollar a year potential market, that's huge. If you could have a generally intelligent system that could do even a fraction of that, that's pretty cool. But in order to get there, it's very hard work and there are lots of easier problems along the way that are not such hard work. And that's where most of the money has been made. So Facebook newsfeed grew Facebook a lot and also created polarization in society. It was part of surveillance capitalism. I'm not sure the world is better off for it, but it certainly made them a lot of money. And it didn't require general intelligence. It just required predicting what stories people

might like based on relatively superficial features. And maybe they could have made it like 7% better by building an artificial general intelligence, but maybe the artificial general intelligence would cost $100 billion and they could do this other thing for $100 million.
And so the big companies tend to pick narrow solutions that do a particular problem well and may not be incentive to really solve the language comprehension problem and the reasoning problem that are at the core of all of this. It'll change eventually, but it's hard.
I think that is a good place to end. So as our final question, what are three books you'd recommend to the audience?
I will start with Steve Pinker's The Language Instinct for two reasons. One is it's the clearest book about what language is really about. And in the context of the conversation we're having, I think part of the problem is that people don't understand the difference between superficially predicting a sequence of words and really comprehending language.
And I think The Language Instinct is very good around that and also around something we didn't talk about so much, which is innateness. We hinted at it. It's great that systems learn things, but some things need to be built in. And Pinker makes a really good argument around that. The second book I'll recommend is Vaclav Smil's How Things Really Work, which I haven't quite finished, but I've just started it. I think it's a great sober look at all the kind of dynamics behind everyday life. I have tried to provide that kind of sober look with respect to AI and I love seeing it in other domains. And then I'll recommend a fiction book, which is The Martian, which is a wonderful page turn, or probably many of your audience have already read it, in particular because it inspired a line in the film version of it, which is, I'm going to need to science the shit out of it. I think we as a society need to science the shit out of a lot of problems and not just have policymakers kind of pull things out of the hat. So that's why I'm going to recommend The Martian.
Gary Marcus, thank you very much.
Thank you.
Thank you for listening to the show. As always, if you want to support us, you can give us a rating in whatever podcast app you're using or send the show to a friend.
Here's a client show that's produced by MfaGaWu and a Galvin Jeff Geld, Roger Karma and Quistin Lin. Backchecking by Michelle Harris, Mary March Locker and Kate St. Clair, virtual music by Isaac Jones, mixing by Jeff Geld, audience strategy by Shannon Busta. Our executive producer is Andy Rostrosser, and special thanks to Pat McCusker.