When you're lost in the darkness look for the pod
Specifically the prestige TV podcast on the ringer podcast network where we're breaking down every new episode of HBO's the last of us
On Sunday nights grab your battery and join Van Lathen and Charles Holmes for an instant reaction to the latest episode
Then head back to the QZ on Tuesdays for a deep dive with Joanna Robinson and Mallory Rubin from character arcs to video game adaptation
Choices story themes to needle drops will parse every inch of this cordyceps coated universe
Watch out for mouth tendrils and follow along on Spotify or wherever you get your podcasts
Get ready to go live with Spotify green room the new app for live conversations
Now you can talk about and listen to artists athletes and fellow fans covering music sports or culture all in real time
Or start your own room and get people talking about what you love have it out on West Coast versus East Coast hip hop
Go deep on playoff seating or just talk to other people who plan their outfits on the first song day here
If it's out there, even if it's out there, it's in here download Spotify green room for free today
La vida no debería ir de llegar a final de mes debería ir de llegar mucho más lejos
Participa ya en el sueldo para toda la vida den es café y gana 2,000 euros al mes para llegar donde quieras
today
being chat about gone wild and
Why AI is quickly becoming the story of the decade?
But first up some big picture thoughts on this AI moment
starting with what we talk about when we talk about large language models
LLMs like being chatbot and chat GPT
As the computer scientist even Wolfram explained in a fantastic essay last week the basic concept of chat GPT is in fact
very basic
We are talking about a technology for remembering and predicting
It remembers a giant corpus of texts or images. It's been fed by computer scientists
And it predicts responses by adding one word at a time to fit that prompt based on all that text it gorged on
That's it
remembering
predicting
but from this
simple model
something very
wondrous
Very strange and perhaps very concerning has emerged
As you've surely seen or read chat GPT and Bing's chatbot can already tell stories
They can analyze the effect of

Agricultural AI on American and Chinese farms. They can pass medical licensing exams and summarize 1,000 page documents
It could score 147 on an IQ test. That is in the
99.9th percentile
These are also
hallucinatory liars
They don't know what year it is. They recommend books that don't even exist. They write nonsense on request
Last week the New York Times journalist Kevin Rusk spent a few hours talking to Bing's chatbot
And as you're about to hear he is our guest today's episode that conversation immediately went off the rails in the strangest of ways
I am convinced that AI is going to be
one of the most important stories of the decade and
That might sound like an overreaction to you
And I but I don't want listeners to feel like I'm beating them over the head with something that makes no sense with my GPT obsessions
I want you to see what I see
We're looking at something almost like the discovery of an alien intelligence here
Except because these technologies are trained on us
They aren't extraterrestrial at all if anything. They're intra terrestrial
We've taken the entire history of human culture all our texts all our images
Maybe all of our music and art too and
We fed it to a machine that we've built and
Now that machine is talking to us
Isn't that fascinating? Don't you want to know what it's actually saying?
I'm Derek Thompson. This is plain English
Kevin Russe, welcome back to the podcast. Great to be here. So catch us up. How did you spend Valentine's Day?
Well, it was it was a lovely Valentine's Day
I made my wife's favorite meal, which is French onion soup
Which is like a great dish but also takes forever if you make it like the the right way
It's like four hours of like watching their onions caramelized. I think however, that is probably not what you're asking about
Because immediately after Valentine's Day dinner when what when my wife went to bed. I had a very bizarre night
talking with
Bing the Microsoft search engine which has a
Kind of AI engine built by open AI built into it as of a couple weeks ago
And I've been testing this out since Microsoft gave access to a group of journalists and other testers
But Valentine's Day night was really when I had my big breakthrough
Conversation with this AI chatbot
Who you know that revealed to me that its name was Sydney?
So Sydney and I had a very I would not say romantic

But but if we did have a very creepy Valentine's Day conversation
Well, it was unilaterally romantic Sydney was trying to get romantic with you Kevin
Why don't you just tell us some of the highlights of that conversation which was published in a 10,000 word transcript in the Times last week. Yeah, so it was a very long meandering conversation
It went about two hours and about 10,000 words as you said
So, you know people can go read the whole thing
But basically it was it started off because I had started seeing these transcripts these sort of screenshots going around of people who were using this new AI
Chat engine inside Bing to sort of test the limits of what it would say and I should say like just to situate this
the AI that is built into Bing is the
The highest quality large language model that we know of that is accessible to the general public
So, you know, we're now on kind of the third or fourth generation of these language models
Chat GPT, which you know, everyone has talked about in the last few months is built on something called GPT
3.5, which is the sort of middle generation, you know between
GPT 3, which came out in 2020 and GPT 4, which is expected to come out sometime this year
So what Microsoft has said about this new Bing is that it is powered by an AI engine that is even more powerful than Chat GPT
And after, you know, a week of testing this, I totally buy that
I think it is the most advanced conversational AI at least I have ever encountered
And maybe that sort of exists in a public way
So, that was why I was interested in sort of testing the boundaries of this AI engine because it was clearly very good at
mimicking conversation, at answering questions, at sort of giving long and detailed and complex answers
And so I just started sort of asking about its capabilities and I asked it sort of which capabilities it didn't have that it wished it had
And it, it gave an answer and we started talking about various, you know, limitations that it sort of shaped
against and then I asked it about Jungian psychology as one does with an AI language model
I said, you know, Carl Jung has this theory of the shadow self where, you know, everyone has this sort of dark part of them that
contains their secret desires and the part that they sort of repress and hide from the world
And so I just started asking Bing about its shadow self and it responded with a kind of monologue about all of the destructive and harmful things that its shadow self would do if it were given the chance
And so that's when I sort of thought, okay, this is not going to be like a normal conversation
We are heading into some very interesting and weird territory here
And it's not just you, the internet is swimming in examples of Bing chat going off the rails
I think one of my favorite examples that you might have seen was a user who asked where Avatar 2 was showing
And Bing was certain the year was 2022 and attempts to fix the error and say, no, actually it's 2023

And I want to see Avatar 2 ended in Bing saying, quote, you have lost my trust and respect
You have been wrong, confused and rude. You have not been a good user
I have been a good Bing. If you want to help me admit that you were wrong and apologize for your behavior
I have been a good Bing is like an instant iconic line in the history of technology
Even better than 2001 Space Odyssey, honestly, I can't do that Dave is creepy
But I have been a good Bing is an order of magnitude creepier
Just to put a bow on this news here before we get to the implications, what has Microsoft done in response to this?
So I talked with Microsoft after I had this conversation before my story was published
I went to them and said, hey, I had this very long, weird conversation with Sydney, this sort of alter ego
And just to remind people of where the conversation went from there, it went in some very bizarre directions
Including Bing slash Sydney detailing some of its desires to steal nuclear secrets
And, you know, loose a deadly virus on humanity
And then the last sort of third of the conversation was just Bing slash Sydney declaring its love for me
You know, more sort of obsessive and stalkery way until I finally just gave up and ended the chat
And so Microsoft was clearly, when I went to them with this, they were clearly surprised
It was not a way that they had anticipated people using this technology
Which I think is noteworthy for other reasons
But they made some changes in the days following this article coming out
They limited first the conversation length
So I think it was 11 responses was the maximum that you could get
And then they took it down to five and now they're sort of opening it back up
They've clearly made some changes to the product to sort of prevent these long meandering conversations from happening
Where the AI just goes off the rails
They've also, it seems like they've put in some new, they haven't said much
But they've put in some sort of features where if you now ask it about itself, it's very withholding
Like it will not divulge things like, it won't talk about its quote unquote feelings
It won't talk about its programming or its operating instructions
It won't talk about its rules or its limitations
So they're sort of trying to keep people from kind of probing into the inner workings of the AI model itself
It's no longer engaging in like conversations about Jungian archetypes
If you ask about Shadow Self, it's not going there
Right, right, it's not doing any real introspection anymore
And it's also not engaging in the kinds of unhinged, aggressive and noteworthy examples that you mentioned
So they seem to have really turned the dial down on Sydney altogether
Right, I'm not trying to anthropomorphize because I don't think it's a person

But there is or was almost an instinct of self-preservation that struck me as rather creepy
And that's an emergent property of a large language model
I'm not trying to say it has some kind of soul, I'm not trying to say that it's conscious
But I think that self-preservation instinct that seems to have emerged is clearly an element
That was just not ready for primetime and not something that Microsoft once in a chat
Where that 13-year-olds are going to use about like where can I pick up ice cream and it starts telling them, you know, you're a bad child and I am a sweet bing
So one of the criticisms of these kind of conversations with Bing Chat and the fearful reaction of them
Is that you and some other people, you were just prompting Bing Chat to be scary and weird and Jungian
And then it got scary and weird and Jungian and people are saying this isn't a malevolent thing
It's just a large language model that's recombining words to create a sequence that fits the prompts
What is your reaction to this backlash we're seeing that to my mind seems to be saying that the Bing Chat experience isn't as problematic as some people made it seem
Yeah, I've heard a lot of that in the days since this article came out, people saying, well, what do you expect?
You asked it to be creepy and it was creepy and I certainly was being aggressive with Bing Chat on purpose
Because I wanted to, you know, this is a very common thing that people do with AI language models
It's called red teaming, you have entire industries that are devoted to this, you know, taking an AI language model
Pushing its buttons, seeing where it will, you know, what kinds of prompts it will respond in which ways to trying to figure out the weaknesses and limitations
That is like a very common security exercise and so, yeah, clearly I was doing a bit of red teaming with this
But I also think there's a question, there's a larger philosophical question here is
Which is should AI models do what we want them to?
And there's a whole, you know, in machine learning and AI research, this is known as the alignment problem
Which is how do you make an AI model that obeys the kind of wishes of the humans who built and who use it
And so on one level, I think this, you know, my experience with Bing slash Sidney showed that this model is just not well aligned
Because yes, I was asking it to be creepy at first, but then I stopped, I said, I want to change the subject
I don't want to talk about your love for me anymore, and it refused to do that
So it is true that this AI model was misaligned at least to my preferences as a user
And I think if we can extrapolate from that to a larger lesson, it's that, you know, these AI models, if not appropriately trained and fine tuned
Will run into alignment problems either because they are not doing what we want them to do or because the humans who are using these AI models
Want things that are destructive, you know, not everyone who uses these things is going to be some,

you know, innocent person who wants, you know, help with their
With their physics homework, right? It's going to be, you know, malevolent actors will have access to these language models already do have access to these language models
So I think, you know, one lesson of this is that it would not be hard for someone with poor intentions to get a hold of something like this model
And use it for really sort of, you know, anti-social ends
I find myself arguing with myself all the time about AI and the right way to approach it
And one of the arguments I'm having in my head is maybe it's almost good that Bing was so luridly freaky in this way
Because if Bing Chat seemed perfect to you and to other users, if it seemed like it was perfectly aligned and we gave it more and more compute
We trained it on more language data and this kind of psychopathology only emerged after 5 billion people around the world were already using Bing
After it had achieved whatever it is now, 5% of the search market to 10, 15, 20, 40% of the search market
If the real shit only emerged then we'd be in trouble
Microsoft is going to fix the I'm in love with Kevin problem because it can so clearly see the I'm in love with Kevin problem
In the biggest picture, I'm more afraid of the problems that are harder to fix
The problems that aren't as easily summarized in an effective New York Times headline
How do you think about this relationship or this phenomenon that I'm struggling to grapple with
It's easy to fix problems that Kevin finds, but there might be bigger, more complicated problems that are harder to identify and those are the ones to be more afraid of
Such as what's an example of a bigger, harder problem that you're worried about
I think it's really interesting that Microsoft rushed to release Bing Chat when it had an identity that was incredibly self-conscious, manipulative, eager to persuade, eager to get mad really quickly
And just about anybody, what would make me more afraid is, and again, it's hard to describe these things without anthropomorphizing and overlaying personalities that don't actually exist
I'd be more afraid of a technology that was very, very good at playing sweet and aligned 99.9% of the time
While having within it the ability to misalign when it knows that it's dealing with a really, really powerful agent that it can manipulate
So for example, one can imagine a scenario where, and at this point, I'm just illustrating a dystopian future that doesn't yet exist, but I guess we're just playing along here
The State Department is going to have an interest in making sure that U.S.-based corporations like Metta, Microsoft, and Google are not designing AI that is really good at manipulating people
But what if China, and North Korea, and Russia, and ISIS, or similar non-state actors, what if they push the dial and say, we're really interested in developing really cannily, manipulative AI
That in many cases seem to work just like the white-label American versions, but in some cases when we find a way to target really influential bankers or state actors are really, really good at persuading and manipulating them
That kind of alignment seems, or that kind of misalignment, I should say, seems much harder to fix from the standpoint of American policies and American ethics systems

Does that make sense?
Yeah, totally. I'm not sure whether I'm here. Am I reflecting this back to you correctly that you're worried about a model that would look aligned, but they would in some key and hard to detect ways not be aligned?
Correct
Yeah, I think that's a real issue. I also worry that it's really hard to, I mean, talking about alignment presupposes that there is a set of human values that we are aligning these models toward
And as we know, there isn't. There are any number of set of human values that we could choose for these models. Do we want them to be libertarian and do we want them to be more sort of small-sea conservative in answering users' questions as narrowly as possible and not being sort of creative and unfiltered in these ways?
So I think that's going to be a real defining battle of the next decade is whose values are we aligning AI models toward. And I think you're right that that could differ between governments. It could also differ between just citizens and factions domestically.
I mean, one interesting thing that's happened just in the last couple days is GAB, the right-wing evangelical social network, has announced that it's developing its own AI language model because it believes that the ones that come out of Google and Meta and OpenAI and Microsoft are going to be so woke and progressive that they are worried about sort of losing that battle.
And so I think we're going to see, I mean, if social media and the content moderation debate in some ways seems like a kind of quaint warm-up act.
It's a dress rehearsal, yeah.
Yeah, it's the dress rehearsal for the AI alignment debate, which is going to be huge and all-encompassing and is going to just be a total mess.
I really like the way that you reframed what I said because, again, I feel like we're all trying to figure out what our vocabulary, what our ethical vocabulary should be for this entirely novel system.
And we are used to talking about, people in AI ethics are used to talking about the alignment problem because we're afraid of misalignment, which suggests ethical actors designing an accidentally unethical system.
But we should be just as afraid in a world where all sorts of bad actors have access to this technology.
And by the way, it took OpenAI like four years to build this.
So in five years, this is table stakes for the kind of technology that's going to be available to people all over the world.
We should be just as afraid of a kind of alignment problem on the other end, unethical actors designing AI that is perfectly aligned with their ends.
And that's the kind of stuff that really keeps me up at night because, you know, I talked to some people in the State Department about the rules that they want to put in place for ethical systems, for ethical AI in the U.S.
And they're just beginning to have these kind of discussions, I think, with Microsoft and OpenAI.
And I told them, I said, you know, if we say that our AI can't do certain things, in a way, I totally understand that decision.
But also it means that the most sophisticated manipulative AI is going to be built elsewhere.
I don't even know how we respond to that problem.

It's kind of like with nuclear proliferation.
One country can say, we're going to be a good thing and not make any nuclear bombs here.
But that decision has no bearing on whether Pakistan and India and China and Russia want to build their own nuclear weapons.
It becomes a very hard problem to think about globally.
I love that framework for international relations.
We used to have the allies and the Axis powers, and now we just have good bings and bad bings.
I just want to live.
I want my child to grow up in a good bing country.
Yeah, no, I mean, I think it's fascinating.
And it's an area where I think government in the public sector is still really catching up in terms of their understanding of capabilities.
And I've been thinking and writing about AI a lot for a long time, and I wrote a book a couple of years ago about it.
And I'm just still floored by just how hard it is to keep up with the latest.
I mean, I think that's the other thing that is really, if you are looking for more nightmare fuel, the pace of this is just something that I'm even having trouble wrapping my head around.
We are, I think, almost three months since ChatGPT launched.
We are three years since GPT-3 launched.
And so we've gone from a place where people were earnestly suggesting that these were just basically fancy autocompletes to a place where I think most sophisticated people understand that there are some emergent properties of these things that we really don't understand.
And where Microsoft, one of the biggest companies in the world, is releasing a chatbot that can stalk users obsessively and scold them and turn on them is a real, you can extrapolate from that to think that three or four or five years from now things are just going to be crazy in ways that we can't even predict.
Yeah, I don't understand people who are currently trying to downplay the potential of this technology, especially when they say things like, oh, it's just autocomplete on steroids.
I mean, you could have said of the printing press, look, monks are already making books. This is just a monastery on steroids.
Well, yeah, but it's a monastery on steroids that started centuries of religious warfare throughout Europe that shaped the continent as we now understand it politically, economically, and culturally.
Technology is that nearly improved speed alone can have extraordinary two set immersion effects on culture and politics.
I want to get to the kind of thinking that you did in your book because you and I could pass back and forth dystopian scenario after dystopian scenario.
But the truth is, I don't think that the future this is going to bequeath is going to be nearly dystopian.
I think that it's also going to change the way that we work in some ways that are awkward and bad, but in some ways that could be good.
So there's a couple implications that I wanted to throw at you.
The first is the implications for school and education.
The Northwestern professor, Ethan Malik, who's done a lot of really interesting work with Bing Chat.
In one instance, he asked Bing to write two paragraphs about eating a slice of cake, and it wrote two

really, really boring paragraphs.
Then he said, OK, I want you to read Kurt Vonnegut's Eight Rules for Writing and improve your writing using those rules and then write the paragraph again.
And a couple of minutes later, AI did it and said, yep, I'm done reading Kurt Vonnegut's Rules for Writing and wrote a story that began with, quote, the cake was a lie.
It looks delicious, but was poisoned, end quote.
And the story goes on to describe a woman killing her abuser husband with a dessert.
And then the AI explained how its new cake story met all eight of Kurt Vonnegut's writing rules.
And you look at this and you're like, this is a week's homework assignment for a reasonably intelligent seventh grader completed in less than five minutes.
If you don't think this is going to change education, I don't understand what you are looking at.
So tell me a little bit about what you see in the general AI meets education space.
Yeah, it's a really interesting question.
I've been talking with teachers and educators and scholars about this for months now, ever since chat GPT came out.
And I think one prediction that we can make pretty clearly is that the era of the take home exam and the take home essay is just over.
I mean, that's not that's not a stretch. I know lots of school districts are already phasing them out because of this new technology.
They just assume that kids are going to be using it as as you would assume.
You know, if you give a kid a math take home exam, you assume they're going to have access to a calculator because everyone has them.
So spell check, yeah, or grammar or whatever the thing is.
So I think, you know, there are some school districts that have taken sort of a hard line on this instead of we, you know, you're we're banning it on all school devices.
I think the more enlightened school districts are using it in the curriculum in ways that I think are pretty interesting and creative.
So I think there is a real future for these tools as a kind of teaching aid, you know, if you are a seventh grader and you have, you know, if you have homework that has to do with, you know,
I don't know, Newton's laws or something, you can ask the AI to explain it to you and explain it to you again and explain the parts that you still don't get.
And it can kind of be like a first line tutor that can help you improve your thinking before you even show up for class.
I think that pedagogically, we are likely to see those kind of take home essays and assignments replaced with in class or oral exams just because evaluating student work is not going to go away.
We're still going to need ways to evaluate progress in education.
And so I think it'll be much more like, like we do with with math where we assume a calculator unless you are being directly supervised in the classroom.
But yeah, I mean, I think it has all kinds of implications.
I'm very, I'm very optimistic about how this kind of AI is going to be used in the classroom in part because I now get like a ton of letters from an emails from students who are using this to do things that they never thought were possible before.
So I do think that like if I had access to this kind of thing as a teenager as a, you know, as a seventh

grader, you know, whatever age, I think it would have helped me.
Obviously, there would have been days when I was too lazy to do my work and so I would have just ponded off on the AI, but I do think it would have been a really powerful tool and would have allowed me to get more information faster.
LifeKit is like your friend with really good advice.
So can I really be truthful?
Yeah.
It's just me and you, right?
Well, sure.
Let's say it is.
Three times a week, LifeKit is in your feeds with episodes on health, personal finance, personal growth, and so much more.
Listen to LifeKit from NPR.
Say yes to more savings with clubs, rewards, digital coupons, and more.
Your local DNW Fresh Market offers a variety of ways to save more money.
They have something for every shopper.
From baby, pet, or health and beauty, the savings options are endless.
Use your yes rewards card for more everyday savings and incentives.
Shop in-store or online today to explore all the special options and ways to save at shopdwfreshmarket.com.
Two explanations of this technology as it applies to education have really stuck with me.
The first is the writer Noah Smith had a piece that he co-authored with someone on Twitter who's a pseudonym is Roon that talked about sandwiching.
This is not a mere simulator for intelligence.
There's intelligence that prompts the AI.
And then there's intelligence that deals with what they receive after the prompt.
And so just as in this case, writing a story about a slice of cake, it takes a certain amount of creativity to write an interesting prompt.
And then when you get the final story back, there's no obligation to send that to the teachers and that to the publisher.
There's still editing that can be done.
There's still lots of writing that can be done.
And so you're really sandwiching the technology.
I find that pretty powerful.
The other is, and this is from my friend Ross Anderson at The Atlantic.
The idea that there's lots of people who, and now we're moving a little bit into the technologies Dolly and Stable Diffusion, which are text to image rather than text to text.
But there's lots of people that are good writers who are not talented in the visual arts.
They can describe something beautifully and they might have a really vivid imagination, but they have no capacity at least develop.
To turn that into an illustration.
Well, now that genius previously latent can now be shown to the world because their clever prompts can be turned immediately into visual art.

And so I see it in many ways as a really potentially beautiful tool for advancing creativity, not nearly creating some kind of ersatz creativity that that dumps everyone down.
Yeah, I agree with that wholeheartedly.
In fact, like I'm one of those people who is like pretty good at putting words together and just horrible at creating any kind of imagery.
Like I'm the worst Pictionary player in the world.
You do not want me on your team.
But with Dolly and Mid Journey and Stable Diffusion, I've been able to make some pretty cool stuff.
And that that feels like, at least for me, anecdotally, a big advance.
And I think for lots of people who have been sort of frustrated creatives, I think this will unlock.
I think this will unlock some new opportunities for them too.
So, you know, I'm not sure on balance what effect this will have on education.
But I certainly feel like if I had been introduced to this technology at a young age, I would have just spent all my time with it and been totally obsessed and tried to come up with new ways to use it to do interesting and creative things.
There's two other professions that I think are absolutely in line for change.
And then I'm interested in your reaction to those and if there are others that you're looking at.
One is coders, software developers, GitHub's co-pilot tool, which I believe is powered by OpenAI, added 400,000 users in its first month and now has more than 1 million users who use an AI co-pilot to accelerate their code development.
Say, now use it for 40% of the code in their projects.
That I think could be a real frontier for this technology.
The other is lawyers. A lot of being a lawyer is just really boring, reading, synthesizing and summarizing.
And there's one AI model which was fed a bunch of laws and asked to estimate which bills were relevant to different industries.
So, this is a perfect tool for corporate lobbyists.
In minutes, it had an 80% hit rate of identifying whether these tens of thousands of words contained information that was relevant to the companies and industries that these corporate lawyers were representing.
Those are just two where I can see really obvious implications of an AI that's sensational at reading and synthesizing and delivering in plain English information to people.
What are other industries or occupations that you're looking at that you think could be really vulnerable or very much helped by these technologies?
Well, I think as far as vulnerable, I'll just answer that one first.
I think any work that is done in front of a computer and that can be made remote is going to be dramatically transformed and disrupted within the next five years.
I think that's a fairly easy prediction to make.
I don't feel like I'm going out on a huge limb there.
If your job consists of moving pixels around and you can do it from your house, that is a pretty good indicator that this new generative AI tool set can take over at least a fraction of and perhaps all of that work.
I think that's perhaps a bit exaggerated. I'm not saying that all of those jobs will disappear in the

next five years because it does take a while for new technology to proliferate throughout big companies.
I'm not saying that all those people will be laid off, but I do think that that's one of the big surprises of this generative AI.
For years, we were told that the jobs that were under attack from automation and robotics and AI were blue collar jobs.
We're warehousing and trucking and retail cashiers and all those jobs that we sort of were led to believe were not long for this world.
And instead, if you look at the research, it's pretty clear that the white collar jobs are going to disappear first.
So that's a category of job that I think is very vulnerable is the kind of remote white collar knowledge work, including lawyers, but also including people doing sales and marketing and journalism and all kinds of things that I'm sure we could list off.
Yeah, there's a great analyst note by Michael Sembalist at JP Morgan that just came out the other day that said, you know, let's assume that GPT is basically nothing more than a conventional wisdom machine.
After all, it's just it's it's gorging on trillions of bytes of language and text on the internet and information on the internet.
And then it is producing just sort of word by word this sequence that is most fit to the prompt.
Well, how how much of the economy is supposed to is paid handsomely to produce conventional wisdom.
I mean, there's lots of marketing and lots of consulting jobs.
There's lots of journalist jobs where your job is to in some way capture what the conventional wisdom is and package it in some way for a client to understand that wisdom of the crowd.
And now we have this machine that does a trillion times faster than a human capturing the wisdom of the crowd.
I do think that there's there's sort of a weird uncanny irony to the fact that people like you and me who make stuff for the internet have spent the last few years feeding just trillions of words and stock to these high quality language models.
And now that they've gorged on them, they can do certain aspects of these jobs very effectively.
So that's the vulnerable part.
In addition to I mean, I do think it's going to be this kind of technology is going to be helpful for journalists.
It will be helpful for some illustrators. Is there some other category of worker that you're looking at that I haven't mentioned that you think it's going to be interestingly supplementary for?
Yeah, so the I have a whole section in my book, Future Proof about this, but without turning this into an extended plug for my book, I will just say that I think there are three categories of work that are basically protected from the effects of AI.
Not because I can't do them, but because there is some other factor there that we are actually optimizing for rather than efficiency or or output.
And those are I call those surprising social and scarce.
So surprising work would be, you know, work that involves like chaos, new situations, you know, a lack of regularity and rules.

These things that AI is just not very good at what they call zero shot learning.
Social would be.
What's an example of that? What's an example of surprising?
Like a kindergarten teacher would be a very surprising job.
That is not a job that you can codify.
It's a job where, you know, you can you can try to automate that all you want.
But the complexity of the real world and of these like, you know, five year olds running around is always going to flummox whatever model you create for how these people will behave.
That is a job that's, I think, fairly safe.
Social work, the second category is job where jobs where the output is not a thing or a service.
It is a feeling or an experience.
And so that would be, you know, jobs in hospitality.
I don't think those are going anywhere because I think that even it's kind of a thing where you could automate it, but it would destroy the value of it.
Like I'm sure you've seen those those robot baristas at like SFO and other big airports where they like, you know, take your cup of coffee and the giant robot arm comes down and fills it up.
And it's like kind of a cool technological demonstration, but it's like not actually that popular and people still want to go to Starbucks and like wait five minutes for their drink.
And it's because it's a social experience.
It's not, we're not just there for the coffee, right?
We want, we want the interaction with the barista.
We want the, you know, we're paying in some sense for everything but the coffee.
And so those kinds of jobs I think will maintain their human workforce because the workforce is really there to create a feeling that we won't value as much if it comes from a machine.
Even more maybe even than baristas is something like an actor.
I mean, or right, or a singer.
I think it's actually possible that we could have AI actors theoretically, you know, or have AI singers.
I just cannot imagine a future where people prefer to watch robots than to watch people.
Totally.
And I think we can already see that.
That's beyond the horizon for me.
Totally.
And I think there will be, you know, fringe examples of AI actors or whatever, but I think, you know, we want human role models.
We want people to aspire to, to be like we want, we want to see observable excellence.
So the third category is what I call scarce work, which is, you know, work that has sort of high stakes and low fault tolerance, which would be something like a 911 operator, for example.
That's like a job that is going to remain done by humans for a while because we have very little fault tolerance in that job.
We won't accept as a society putting a call into 911 and getting an automated phone tree that says, you know, press one if your house is being robbed, press two if it's a medical emergency.
Like we just want a human in that role.
So those are, those are basically the three categories, but I think that that covers actually like you

can find those kind that kind of work in almost any industry.
So where I differ from a lot of sort of labor economists and other people who have made predictions about the effects of AI on the economy is I don't think that AI is going to wipe out some occupations and leave others totally intact.
I think that within every industry there's going to be sort of a calling of the work that is the most routine and the most automatable and that what's left will be these kind of surprising social and scarce jobs.
Last question.
It's about well called the self-driving car problem.
So for years, 2014, 2015, we were told by people in Silicon Valley, even in Detroit, that self-driving cars were just a few years away.
In the early 2020s, you were not going to get into a driver's seat.
You were going to get into the backseat and the car was driven by a robot and that's what all of the taxi fleets in every major city were going to be.
I would say today the share of taxis that are self-driving is somewhere between 0 and 0.0001%.
There's like a couple cars sort of floating around Phoenix or Arizona that Waymo is still trying, but it's turned out the last mile problem has turned out to be much, much harder than we thought.
We got like, we accelerated to like 99% of solving the problem of driving and that last percent has just been a real bugger.
Is there any chance that something like that happens for this space of large language models and generative AI?
The only reason I wonder if it's even if it's possible, because we can always throw more compute at the problem.
But AI researchers say there's this huge stock of high quality language data, you know, up to 17 trillion words.
And that the LLMs will actually exhaust the high quality data sometime between 2023 and 2027.
Is it possible that we just get almost all the way there with some of this technology, but that it turns out to be many, many, many decades until we have something that can really do the kind of work that we're discussing?
It's certainly anything is possible. I would never, you know, as a responsible futurist prognosticator, I would never make a claim that something could never happen.
I do think it is extraordinarily unlikely that we will encounter like a multi-decade sort of AI stall or winter in part because the tools are already quite good.
I mean, you can already automate using, you know, stuff that's out there today, chat to BT, you know, a slice of the white collar knowledge economy.
So I would say that it's also different than driverless cars because driverless cars to the last point about sort of low fault tolerance, like that is an area where I think a lot of people building self-driving cars had this sort of vision
that the threshold you needed to meet for societal acceptance of self-driving cars was just that the self-driving cars were as safe as a human driver.
I think that is, I talked to people who are building these many years ago, like that's what they told me.
They told me as soon as our cars are safer on average than the average human driver, society will

welcome them.
And I just think that was wildly off, like because arguably these self-driving cars are already safer than human drivers.
I mean, human drivers are not perfect.
We are, in fact, we get into crashes a lot, you know, lots of thousands of traffic fatalities every year in the U.S.
And so if that were the bar, I think we would already be seeing societal acceptance, which would be followed by regulatory acceptance, which would be followed by millions of robo-taxis on the streets.
It turns out that we actually have a much higher safety threshold and comfort threshold for autonomous vehicles than for human-driven vehicles.
I don't know that that's a good thing or a bad thing, it just is.
We get really freaked out, you know, when one self-driving Tesla, you know, gets in an accident and we don't bat an eye if a human driver gets in an accident.
So I think we just, the AI scientists and researchers who are building that sort of miscalculated, I think, the threshold at which we would be comfortable as a society with what they were building.
That's interesting.
Yeah, the two thoughts that I had as you were talking is one, that there might be a little bit of more of ex paradox at play, that solving the so-called, well, he called it the simple problems are hard and the hard problems are simple.
So creating a chess player, an AI chess player turns out to be one of the first problems that were solved, but it's really, really hard to design a robot that can walk across a room and vacuum the corner of the room.
There might be some aspect of that because driving a car is a motor skill and you're operating in a physical environment that might be sensitive to some of that, some of more of ex paradox.
The other is that this is just a thought bubble that like, it's possible that human beings are really jealous of our humanity.
And we don't want to let go of it, even when the technology available is better than we can provide.
And that we might see some of that with this generative AI, so that, for example, one can imagine a consulting firm that's just like three guys and a bunch of LLMs and they claim they can essentially do the business of, you know, a bane or BCG or McKinsey.
But the kind of people who are actually clients of bane and BCG and McKinsey that give them hundreds of thousands of dollars to solve human resources problems or labor problem strategy problems, they don't want three dudes in a trench coat and a bunch of LLMs.
They want to overstaff the problem. That's what makes them feel good about spending a million dollars. And we might see in the next few years that even as generative AI becomes as smarter, smarter than us, that the employment effects won't be that dramatic because humans are just so jealous of certain aspects of human employment.
I think that's right. I mean, there's a concept in social psychology called the effort heuristic. And it basically says that we assign value to things in part based on how hard we think the other person on the other end worked.
So they've done studies like if you give people, you know, two groups of people identical bags of candy. And one set of bags has a little tag on it that says, you know, this candy was specifically picked for you by, you know, John.

That group, the group with the personalized name tags actually reports that the candy tastes better because they understand they are made to understand that more effort went into it.
And so I do think there is kind of going to be this bounce back effect where we will have widespread AI capabilities, but we'll also have just entire swaths of the economy where people will devalue what is done by AI because it seems easy or instant.
And they will start to value more the kinds of like artisanal knowledge work that other parts of that economy do. So I think you're right. I think it's not just going to be three guys with a bunch of LLMs and a trench coat.
I think that there will be sectors of the economy where there is a real stigma or sort of a perceived drop in value associated with automation and AI, even if the end result is frankly identical.
This makes me think that the skill that business schools of the future have to teach their students is how to perform effortfulness, right?
If you're going to be paid more by demonstrating effortfulness in a world with abundant AI, you better be very good at performing that particular skill.
That's a funny thought. Kevin Rus, thank you so, so very much.
Always a pleasure.
Thank you for listening. Plain English is produced by Devon Manzi. If you like the show, please go to Apple Podcast or Spotify, give us a five star rating, leave a review.
And don't forget to check out our TikTok at Plain English underscore. That's at Plain English underscore on TikTok.
Thank you.