Hi, I'm Erica Ramirez, founder of Ili and host of What About Your Friends, a brand new show on the Ringer podcast network dedicated to the many lives of friendship and how it's portrayed in pop culture.
Every Wednesday on the Ringer Dish Feed, I'll be talking with my best friend, Steven Othello, and your favorites from within the Ringer and beyond about friendships on TV, in movies, pop culture, and our real lives.
So join me every Wednesday on the Ringer Dish Feed where we try to answer the question, the LCS, back in the day, what about your friends?
Today's episode is about AI and the release of GPT-4.
But in a deeper way, it is about the spookiness of feeling like you're standing at the edge of some phenomenon that you do not understand at all.
And today's guest is a real talent at unpacking the spookiness of new technology, that's the Atlantic's Charlie Warzel.
Now, if you don't know a lot about GPT, if you've been kind of subtly ignoring the story, you might think, wait, I think I heard about this.
Is this the thing that students are using to cheat on tests?
Oh, is this a technology that powers Bing, that thing that weirdly flirted with a bunch of journalists?
Isn't this just like a weird toy?
And the answer is yes.
It is weird, and it is kind of a toy.
Last week, GPT-4 came out, this is the latest version of the large language model technology made famous by Chat GPT.
You know how it works, you prompt it, it talks back to you, it predicts text and sequences, this is the standard way of explaining the technology.
But I've been experimenting with GPT-4 for the last few days, reading about it rather obsessively, and I want to talk about three of what I consider its most important implications.
First, it is an ace student.
The previous GPT model tried to take the uniform bar exam and scored in the 10th percentile, that is a failing grade.
GPT-4, the latest version, scored in the 90th percentile.
It scored in the 93rd percentile on the SAT reading and writing test, the 88th percentile on the full LSAT, it's gotten a five on several AP tests.
Now, some people are waving away these accomplishments by saying, well, I could score a five on AP
Bio2 if I just looked everything up on the internet.
But this technology is not looking things up online.
It is not rapid-fire Googling answers.
This is a pre-trained technology, pre-trained being the P in Chat GPT.
It's using what passes for artificial reasoning based on a large amount of data to solve new test problems that have never been published online, and in many cases it is doing better than most human beings already.
Second, it is kind of like a Star Trek replicator for content, a hyper-speed writer and computer

programmer.
It can code in a pinch, it can spin up websites based on simple illustrations, it can solve programming challenges in seconds.
Now, Charlie and I are going to talk in a second about how this tool might replace certain tasks in the economy, might supplement certain tasks in the economy.
But for now, let's just imagine a very basic, very prosaic application.
Parents can instantly conjure original children's books for their kids.
Here's a scenario.
Your son, who loves alligators, five years old, comes home in tears after being bullied at school.
You go to Chat GPT.
Write me a 10-minute rhyming story about a young boy who overcomes his bully thanks to his magic stuffed alligator.
You will get that book in minutes.
If you want illustrations, you'll get those in two minutes from Dahle or mid-journey.
This is astonishing.
Third, in the wrong hands, this will be a terrible nuisance, and that is true even if you don't believe the most apocalyptic predictions for this technology that we're going to get into in a few minutes.
One of the concerns for AI safety researchers is that AI will be able to steal money that it can use to bribe humans to commit atrocities using us as like meat puppets of an artificial terrorist network.
You might have heard that previous sentence and gone like, wait, that is an absurd prediction.
Because AI can't bribe people to do anything, except maybe it already has.
Open AI released a document listing the ways it has trained GPT for for safety.
Before the final guardrails were installed, Chat GPT got a task rabbit to solve a captcha, one of those visual image security tools that says like, click every image with a bicycle to prove you're not a robot.
Well, Chat GPT is a robot, but still it asked a task rabbit to solve a captcha.
The worker responded skeptically and asked GPT if it was talking to a robot.
The computer made up an excuse.
It lied.
It told the worker quote, no, I am not a robot.
I have a vision impairment that makes it hard for me to see the images.
That's why I need the two captcha surface.
The human then provided the results, proving to be a very excellent meat puppet for this robot intelligence.
Now it might sound like I just gave you two good use cases for GPT-4 and one bad use case.
But even the examples I just gave are actually more ambiguous than they might initially appear.
For example, number one, I said GPT-4 was an A student.
And maybe we can use its incredible inference skills to raise the ceiling of human intelligence.
Or maybe not, maybe kids will just use it to cheat on tests, which would actually lower the ceiling of most individuals intelligence, like 10 years from now do a podcast on how

some super smart AI is ironically making all of our kids dumber.
Second, I said GPT was a bottomless font of content, right?
Maybe it helps parents and kids and creators come up with ever more amazing artistic ideas and we end up with the best pop songs ever, the best movies ever.
Or maybe we just don't.
Maybe we just get more of everything, more shit, more cheap, sub replacement level nonsense.
Finally, I said GPT-4 could order people around.
That sounds pretty bad.
But ordering people around is a big part of the economy.
That's what managers do.
What if there's some weird future where middle manager AIs are so good at their jobs that corporate productivity skyrockets and the white collar work becomes a four-day week?
Again, what are we looking at when we look at this technology?
I've told this story before, but it really captures my ambivalence and my ambiguity about this whole space.
Imagine you saw a picture of an embryo at 10 days, is growing almost exponentially.
You can start to make out possible organs, limbs.
Someone asks you to predict what kind of an animal this is.
Is it a frog?
Is it a dog?
Is it a wooly mammoth, a human being?
Is it none of those things?
Is the species we've never classified before?
Is it an alien?
All you've got are three data points, 10 days, exponential growth, new living thing.
All you know is that it's larval and it might become anything.
I don't think AI is alive.
I don't think it's conscious.
But I do think it's larval and I do think it might become anything.
I'm Derek Thompson.
This is plain English.
Charlie Warzel, welcome to the podcast.
Thanks for having me again.
I wanted to bring you on because I consider you a bard of uncertainty when it comes to technology.
You are very, very good at diving into deep murky waters all the way to the bottom, seeing what's going on at the ocean floor and coming back to the surface and being like, holy shit.
There's some weird shit down there.
You're very good at explaining the quality of weirdness that you observe in all of these spooky corners of technology.
AI, I consider a very spooky corner of technology.
I do think that lots of conversations about AI that I hear on other podcasts can immediately go into the stratosphere of speculation very, very quickly.

The truth is, we are headed toward the stratosphere of speculation in just a few minutes.
Before we hit blast off, I want to start by anchoring the conversation to things that are actually happening, actual news.
GPT-4 is out, the fourth generation of this technology from OpenAI.
I am using it.
I have forked over the 20 bucks a month to get access to the chat GPT that's powered by this tech.
Give me a sense of who else is using this technology right now.
I think right now, there's many different camps.
There's the true sickos like you and I who have forked over the money because we just need to experience this now and get our bearings and we're going to write about it.
There's that exploratory crew.
Then there's the people who may or may not know that they're using it.
That's the people who want to use Bing's new chat bot.
That is infused with, as we now know, it was speculated, but we now know that's GPT-4 or an early version of GPT-4.
There's all these different, think of it as software updates.
On that sense, Microsoft just rolled that out to anyone who wants to use it.
You could say that millions of people are using that today to do that prosaic search chat bot thing that we've been talking about for a while now.
Then there's the enterprise group, which I think is a super fascinating use case.
This is the world, I think, that's going to actually, this is how Uncle Steve or Aunt Molly or whatever is going to start to encounter the technology.
These partnerships through OpenAI has a partnership with the consultancy firm, Bain, to work with clients like Coca-Cola, these big, huge companies.
They released an API integration, which is essentially allowing different programs to access the tool or different developers.
We're seeing Slack is developing one of those to respond to messages or summarize big, long threads in very concise bits.
Salesforce has that for their customer management stuff.
They're going to roll that out, and Salesforce is used by tens of millions of people to do really boring stuff across businesses everywhere.
Then you have the announcements this week from both Google and Microsoft of they're going to put this generative AI tool stuff inside all their workplace clients.
That's docs, calendar, Gmail, slides, whatever.
That's going to be able to do and automate a bunch of that different stuff in the way that you currently have autocorrect for your Gmail.
Really, it's hard to know how many people are using this tool and in what way.
There's the purest version, which is the, I pay $20 a month and I'm just going to experiment my face off, but I do think that there's a number of people who are encountering this in a very organic way just through their jobs or at least will very soon.
Charlie, there's all these ways that people are using this and showing off their usage online on Twitter.

Give me an example of what you consider one of the most clever applications of this technology.
There's one that I saw yesterday and my journalist brain is like, this is like, I love stunt journalism.
I was like, this is perfect.
It's basically, I think you probably saw it somewhere along the line if you've been looking at this stuff, but somebody basically said, I want to take $100 and start a business and I want to have GPT-4 make decisions for me to try to turn that into as much money as possible without doing anything illegal and just sort of refine the steps along the way.
It's like, what kind of business and I think they decided on environmentally friendly products like silverware and weird stuff like that for camping, but then like, okay, so what will the website look like?
What will the logo look like?
And then feed that into a stable diffusion or mid-journey prompt and get something out and refine it.
It's really fascinating.
It's really cool to sort of see, I think we're used to, and I noticed this with Bing search.
We're so used to with machines to be like, give me one discreet answer and not to have the computer or the machine reason at all or make multiple inferences, but like the genius of AI-assisted search is you can say, how do I get this IKEA bed to fit or can I get this IKEA bed to fit in the back of my Ford Saturn or whatever, Ford Fiesta and it will like go and do all the different calculations and look up all the different stuff.
So I think that's where I and like a lot of normies need to start to like change our brains, right?
It's like, how can we get this thing to start thinking a little bit on our behalf or at least, you know, taking steps and making connections because that's what this technology ultimately does is just makes lots of inferences, right or wrong, as opposed to like need an answer, ask for an answer, get an answer, transaction over.
There's two really, really interesting things that you've made me think of.
The first is that you're so right that I have felt consistently jealous being online, seeing other people come up with just these ingenious ways of using this technology and it's, it makes me think or perhaps it just reminds me that technology always unlocks previously unlatent, previously latent skills, right?
Like the fact that some people are incredible drivers for NASCAR is a skill that had to be unlocked by the invention of the car.
The fact that some people are incredible at GPT is something that maybe just, you know, I'm an amazing prompter.
That's a kind of creativity that I don't know what I would have called it before we had this technology.
Like I think you're a very creative person.
I'd like to think I'm a creative person, but I go online and I'm like, wow, I am not remotely good enough at this particular skill that has, whose door has been opened by the invention of this new technology.
I think what we would have called that skill and why a certain subset of people are great

at it is we would have called it engineering, right?
Because what engineers do is like when you're coding a lot of the times is they're like getting machines to, they're like programming a series of steps, right?
Or like algorithms or like guidelines for a machine to do certain things and take all these different inferences.
And that's sort of what this is, right?
So it's sort of like, whereas like the age of Google was the age of like a good librarian or like researcher was dominant, like the age of GPT, whatever is the age of generative AI is sort of like the programming mindset is really dominant more than the librarian or researcher one.
It's like, how do I sort of give parameters to something and allow it to do a lot of work on my behalf?
So yeah, it's a totally different mindset.
And I'm okay at saying, at least right now, that I'm not as good at it as a lot of people.
The other thing your comment made me think of is that it kind of feels like a boring place to start, like what other enterprises are using this now, like Bain and Microsoft Enterprise.
And the reason this is such an interesting phenomenon to me is the way it fits into the history of technology, the way that most game changing products have typically launched is that they started off expensive and rare.
And then over time, they got common and cheap, right?
That was electricity.
That was the personal computer.
That was the iPhone.
They start as luxury products and they become common products.
This is the opposite.
Like right out of the gate, these tools are cheap for many people.
They're utterly free and they're absolutely ubiquitous.
We have maybe hundreds of millions of people around the world that have used this technology at some point in time.
They're red teaming it.
They're showing how it's wondrous.
They're showing how it's scary.
In your conversations with experts, and I really now want to move into this wonderful article that you just published, what did people say would be some of the most interesting implications of the fact that this technology isn't, it's being grown in one million Petri dishes all over the world at the same time?
Yeah.
I mean, there's not a great, there isn't a huge precedent for it.
And what's actually interesting in that comparison is that like the cost thing is going to probably scale in the opposite direction, sort of in the macro sense, right?
Like the more powerful these tools get, like a number that I heard talking to some people is like a GPT six or seven supercomputer could theoretically cost on the order of $100 billion.

So it's like, we don't have, just the thing that powers it will cost more than most large companies that exist in the world, right?
So I think what's interesting about the usage like this is that it's going to probably follow to some degree that sort of the best analogy I got for like the moment of where we might be over the next couple of like months or years is really like a little bit of the 1990s, late 1990s consumer internet.
And like the idea that the phrase that this person used who works at an AI startup was just like that these generative AI tools would just be nodes that sit on top of things, right?
And a way to think about that is like connecting anything to the internet, any service, right?
You connect a service to the internet and it's not like the service is completely unrecognizable, right?
Certain things change about it, certain things like certain elements of how that service or company or whatever operates, if you like sell tickets or whatever, like your box office in meat space is probably going to be a lot less important, right?
And so jobs will change in that kind of way.
It doesn't necessarily mean that they will get eliminated, but they get fundamentally changed, right?
Or like, you know, I think about all these tools the way that Microsoft and Google are putting them out.
It's like, it's not maybe that your job if you do a lot of like rote communication every day is going to go away.
It's like you're going to be sort of a middle manager of AI tools, right?
You're going to be really good at delegating tasks to them, but also, you know, making sure that you can evaluate like an editor would do with our work that like they're staying, you know, they understand the assignment and they're staying within the guardrails and refining and stuff like that.
So there's this way in which, you know, things are going to get weirder and they're going to change probably, but not sort of in not always in the like apocalyptic way of just like, you know, your profession is eliminated next week because this bot can do whatever it does.
And so I think like that's sort of where we're at, it's almost what makes me exhausted about thinking and covering, you know, this is that it affects like everything.
It is truly in that same way, like if you were to say, hey, like, you know, to some a reporter in 1997, like you're responsible for all internet, like everything that happens on the internet, that's you, like that job becomes impossible in a matter of seconds.
And I think that's the same with like AI, right?
Like we're going to have to, it's just going to be like the connective tissue of a lot of what everyone does and how everyone works.
So that's a little off how I'm thinking about it.
And one of the phrases that another person I interviewed used was that like, they feel like AI is similar to an invasive species in that way, right?
Like this thing kind of comes into an environment and it really dominates and it really takes over.

It doesn't mean it kills absolutely everything around it, but it changes the ecosystem, right? Certain things thrive, certain things are less important or have less resources as a result.
And I think that's a good way to think about it.
Like it sort of fits for me the exact balance of like, I'm a little scared, but also like, okay, there's something evolutionary about it in the sense of certain things win, certain things lose all the time.
I thought that metaphor was really powerful.
And the reason that it clicked for me is that it captures this critical fact of speed.
Like what an invasive species does is not just slowly over centuries colonize an ecosystem.
By sheer dint of its invasiveness, it takes over very quickly and we're looking at a technology that has gone from no one in the world was really talking about, I mean, no, people in AI of course are talking about it, but no one, no, no real normies were talking about GPT five years ago.
And we went from chat GPT to GPT four in a matter of six months.
I mean, this is such a null phenomenon.
I remember this podcast launched about a year and a half ago, our mutual acquaintance, Kevin Roos was on the first episode, name of the episode was the future is going to be weird as hell.
Guess what technologies we used to illustrate the fact that the future is going to be weird as hell.
NFTs and the metaverse, this was not on our radar at all.
I had a feeling that tech was moving in a berserk direction, but this was not on the radar at all.
And the speed with which it's come on the radar has been extraordinary, the speed with which it has proliferated again, 200 million users in six months.
That is invasive species shit.
And the fact that we're all experimenting with at the same time that GPT on any one individual's computer is its own Petri dish, Petri dish suggests that the implications I think could become very weird, very interesting quickly.
Do you mind if we get into some of the conversations you had with some experts about how exactly it would get weird and positive and negative ways?
I want to start with the observation that typically when journalists like you and me reach out to an expert in field, a domain, it's COVID maybe.
We talked to someone about aerosolized viral spread.
The expert typically says, I have a very clear theory of how this virus works.
I'm giving you facts.
These facts come from decades of research.
I feel very strongly about them.
I got the sense from your interviews that every expert you talked to was like, I have no fucking idea what's going to happen.
I spent decades on this subject and I am at a loss for whether or not we are building the most miraculous machine in the world or dooming ourselves.

You had this amazing study of, this is Melanie Mitchell at the Santa Fe Institute, a survey of 480 natural language researchers where they were asked whether, given of data and computational services, could these technologies understand natural language in a nontrivial sense?
They were divided 51% to 49%.
No one knows how these things think and no one knows what's going to come of them.
Let's take just one example here.
You had a conversation with Eric Schmidt, former Google CEO.
He wrote a book with Henry Kissinger about AI and the future of humanity.
What did Schmidt tell you?
This was interesting.
He was kind of actually like, he just wanted to talk, I think he was doing like a little bit of like a press tour, just about sort of, I think the book has, it came out, I want to say like a year and a half ago and it has like renewed relevance.
Anyway, there was a lot of him just trying to explain concepts that I was a little bit familiar with already, but then we got to the, I got to the sort of like, how do you feel about, because in his book, they talk about a lot of nightmare scenarios.
I said, how do you feel about the possibility of a technology like this being unleashed on society the way that it is, given that you and so many people who are building it or excited about it, in the same breath with this is going to, whatever, change the amount of intelligence in the world, this is going to give sort of democratized, like rational thinking or like whatever, whatever the heck way they want to talk about it, in the same breath they're like, and it could be a civilization level extinction event, maybe, and I was like, how do you like balance those things, right?
And he gave this long example of AI powered tools and social networks and like amplification platforms that basically can take anyone's message and make it stickier, make it more viral, make it like punch it up essentially and give it sort of understand, in this way, these networks or these programs would understand what that particular network or an audience wants, right?
Like how to optimize for clicks or shares or engagement, and it would like punch those things up and it wouldn't know like, this is good or this is bad.
It would just say, we're trying to serve you, the creator.
And so the example he gave was someone like, I don't know, like a terrible racist who was just trying to troll, who's trying to do something awful, and that these things would just like naturally create and make this stuff even worse, that would be a feature, not a bug.
And I'm listening to him describe it and I said, well, that's terrible, is that worth, is that risk, just that kind of prosaic risk in one's circumstance at scale, like is that worth what we would get, like the benefit of that knowledge of that, you know, automation and sort of that, you know, expanded, you know, computer consciousness.
And his answer was, hell yeah.
And he couldn't really provide me like a justification for that, right?
He would say like all the big problems in life right now, climate change, you know, speech issues, et cetera, they're all like, we need smarter people and these tools are

going to make everyone smarter.
And I just think that is, it's not that I like fully disagree that those, that could happen, but I just think that it's not a very convincing or imaginative thing.
And this is what I find difficult talking to real like boosters of the technology is they're really imaginative about the downsides.
Like we're talking about, you know, Skynet levels of like computers become sentient and then, you know, try to kill us all levels of imagination, followed by sort of not a lot of imagination about the upside.
It's just sort of like, I mean, to use a phrase that you're familiar with, like abundance, right?
It's just like, it's going to create that and it's kind of hard to understand where we're going to get there.
So that's just a tension that I see and there's not a great explanation.
And then the other thing he said obviously is he's trying to raise awareness because he wants the right people to be building these tools and to learn the lessons of the social media era.
But as I told him, I'm not so sure we've learned all the lessons of the social media era yet because we're kind of just moved on.
We've kind of just onto the next technology.
And there's a huge can of worms with all of this, whether it's Sam Altman who runs Open AI or the people at Meta building large language models or Google building large language models.
It's who, who are we trusting?
Like how are we making these decisions about who we trust to build this stuff if the worst case scenario is they become intelligent, like truly intelligent and kill us all?
Given that if we stop doing anything in this space, if Open AI is shut down and Microsoft stops doing all their AI research and Meta stops doing their AI research and Alphabet and Google stop doing their AI research, et cetera, if we shut it all down, that is nothing to do with China.
It is nothing to do with Russia.
It is nothing to do with some non-state terrorist actor who can get technology that is developed by China and Russia, et cetera.
Given that reality that we just don't have, the federal government doesn't have, the U.S. federal government does not have a monopoly on the question, should we proceed with AI research in the world?
Have you spoken to people, are you persuaded by the case that we should just stop?
I actually don't believe that we can.
And that's what's so interesting to this because if you look a little bit now at, and I'm going to get the real technical specifics on this wrong because I am, my job is to talk to people about this stuff and try to convey, as you said, all the weirdness and uncertainty.
The technical aspects of this whole world are dizzying to me and that's why I try to rely on them.
But from what I can gather, there are already ways that these models are able, the true instance of the model, that people are able to run them locally on really good computers,

like a high-end gaming computer.
And that those models are going to, that's only going to become more and more a thing going forward.
And there's this whole idea of the open source idea of large language models and people who are building them.
And now there's a bit of a fight, almost, or a disagreement between certain companies about open AI, which open is in the name, is now saying, we're seeing it's probably a pretty bad idea to make these things open source because of the ability to abuse them or for them to fall in the hands of the wrong people.
But some of those things already exist, the blueprint already exists.
And I just don't think you can put the toothpaste back in the tube on a lot of this.
Now maybe the most powerful versions, the ones that require hundreds of millions of dollars, maybe we will see, and I'm hearing from certain people that there is a lot of buzz in the federal government right now around, we probably need to really figure out who's allowed to have how much computational power and access to the funds and the resources, and that kind of regulation around this stuff.
But the idea that the tools we have right now aren't going to be able to be spun up by people in a very short amount of time to use as they see fit, that's happening is going to happen.
We're not going to be able to take that out.
And that's a whole different ball of wax, can of worms, whatever, use your analogy than talking about US versus China or US versus terrorist state actor getting licensing stuff from China or whatever.
And that adds a whole another level to this, and it's something, we don't have to go there now, but it's something like I allude to at the end of my piece when it's talking about like, there's some really real questions here about like, should we build this?
Because we should always strive for technological progress.
Should we build this because it's a security imperative to build this?
Should we not build this because it's a security imperative to build it?
These are really interesting questions, and we're only in like hour one of really trying to suss them out.
So you're looking for someone who leads you to your best performance?
The talk is not about online dating, but about online business.
Ideal partner, Shopify.
Ten thousand German companies trust the revolutionary platform that covers all sales channels, personal
POS, social media, e-commerce.
Test Shopify free of charge and bring your business idea successful in the world.
The writer Stephen Johnson has a long piece in the New York Times magazine about Thomas Midgley, who is a brilliant inventor in the first half of the 20th century.
He worked for General Motors for a while solving problems like engine knock.
And he played a huge part in inventing two of the most toxic, most horrible chemicals invented in the 20th century, chlorofluorocarbons, CFCs, and freon.

And it's a really interesting story about how invention can sometimes create monsters.
The problem of CFCs, which were burning through the ozone, had to be solved by a group of scientists and corporations and politicians that came together in the 1980s to do something called the Montreal Protocol, which essentially banned CFCs.
And since then, the ozone has recovered, especially over Australia and New Zealand.
In a similar way, I am worried that we are creating a monster that we won't necessarily know how to stop unless we all come together as a world.
There was a survey that Ezra Klein noted in his column last week, a 2022 survey very recent of AI experts that asked them, what probability do you put on the human inability to control future advanced AI systems from causing human extinction or similarly permanent and severe disempowerment of the human race?
The median reply was 10%.
One in 10 people working on AI think it is possible if not probable these systems will cause human extinction.
That is crazy.
What is your reaction to the fact that this industry is populated by people who think they're working on a doomsday device?
And it goes further than that too, right?
If you look at Sam Altman, one of the founders of open AI, wrote a blog post at the end of February that actually didn't get a ton of traction, I thought, given what it's talking about, but it's working towards an artificial general intelligence, which is the goal of open AI, which is a truly intelligent, conscious, almost entity.
And it basically lays out that exact dichotomy that I was talking about with Schmidt, which is this could be the greatest prosperity generator in human history.
This could essentially unlock unforeseeable amounts of things at a level so much greater than modern personal computing in the internet.
And obviously throws in and the risk is potentially civilizational, if not, some smaller, very concerning security risks, whatever.
And every conversation to a person that I had, I'm talking about engineers building these tools, researchers, like boosters, skeptics, safety experts, truly everyone, they all veer into this territory, right, this like super late night dorm room or like philosophy class territory.
And the thing, I really couldn't find any other parallel with other technologies, certainly not like the social media ones that I have covered for more than a decade now.
I think there are some things maybe like genetic modification, but the thing that I just kept going back to was everything that I've ever read about the Manhattan Project.
It's nukes, right?
Exactly.
And it just feels like when you read about, you know, the hand wringing that was going on in Los Alamos over this stuff and the, you know, like, I'm not going to get the name, but there are people that quit the, you know, quit the project, like high level people because they just said, like, I actually can't, I just can't get on board with it.
And what's interesting about that to me, what I noticed at the end of the piece is like,

they knew what they were building.
Like it was very clear, like we are building, we're, you know, we're splitting the atom.
It's going to create this, you know, weaponized massive explosion that we can control.
It's going to kill a lot of people and it's a, you know, it's a defense thing.
They knew exactly what it was.
When you talk to people in this field who are building the products or who have, you know, can see the guts of them or understand them because they have, you know, math and physics and computer science degrees, they will say, we do not know how we know how these things are trained.
We know how they're refined.
We know how they're weighted.
We do not know how they come to their conclusions, their inferences in the same way that like you can watch on an MRI, a part of the brain light up, but you don't really know how the neuron, you know, fired and gave that exact thing.
There is just a level of, you know, quote unquote, like mysticism around it, just simply because we don't have the ability to understand it.
So in a sense, we are building something in the same way that, you know, they were building that, you know, those nukes, but they knew what was going on and here we just kind of don't know, right?
Like open AI's whole reason for being is basically this is such a dangerous experiment.
We want to do it with the most, you know, altruistic, benevolent values possible, and, you know, with transparency, even though I'd argue there's some issues there, but it's wild.
It truly is like, I've done a lot of technology reporting in my life and I've never had so many like, I'm just like up at two in the morning staring at the ceiling being like, what are we doing?
I love the analogy to the Manhattan Project and it might be even one degree weirder than you portrayed because it's not just the scientists in Los Alamos designing a nuclear bomb that they thought might end the war.
It's also imagined if those same scientists knew they were designing the core energy source of a nuclear power plant.
And so they were thinking we might have in our hands the secret to clean energy forever for the future of the human race, but also we might be launching a bomb and this is something I believe they thought.
There was some fear about this that will incinerate, it will ignite the nitrogen in the air and cause the explosion of the world.
This might be clean energy for the history of the planet or it might incinerate the world.
Like those options could have all been on the table if they had perfect insight into the 1950s progress with Adams for Peace and Eisenhower's nuclear energy investments.
It's mind-boggling to think about all this being on the table.
Speaking of all this being on the table, there is one doomsday scenario that has gotten way more attention than it was any other.
This was from an Ellie Eiser Yadkowski interview in a podcast where he told a story about how he imagined AGI could contribute to the death of the human species.

And I think it's worthwhile to just retrace this story just so that we can talk about it a little bit and just share our feelings about how we're making sense of this particular prediction.
Do you want to briefly summarize this story?
Yeah.
I'm going to be drawing off of a hacker news thread that's summarizing this because we were talking off the air about this and one thing about the AI community, especially those who are sort of the thinking brains in a tank around it, is that they're very verbose or just like loquacious, very long podcasts, very long blog posts, things like that.
So I'm relying a little on this, but basically the theory is that we're going to spend a lot of venture capital money, we're going to put it in AI, and most of it actually will go to waste, but a small part will level the technology up to the point where the AI will be able to write another AI, right?
And then that AI will write another AI, and that one, and that one, and that one, and you sort of get that like replication, almost like, you know, human evolution where like things start to change, unexpected things start to happen till you get to the point where an AI will be smart enough to announce that it's concluded that atoms inside human bodies could be repurposed for something else that it has decided it's better, right?
That human beings, their energy, they're like, there's something wasteful there.
So what it will do then is it will basically try to design a plan the way that any, you know, like foreign power or, you know, basically like a terrorist organization would, right?
And so his example is that the AI will send an email to a human in power with like specific instructions on how to make a bio weapon, right?
The AI will also possibly like hack into and break into a bank and get access or, you know, run some kind of scam, get access to a large pool of money, and it will, you know, it will pay people, human beings to, you know, follow these instructions to make a weapon, basically just to do its bidding.
And, you know, these people might not even know because the AI is so crafty and clever that they're talking, you know, to a sentient, you know, computer network.
It will just, it will sound like it's coming from an organization or something like that, like another person.
Anyway, the other idea too is that, you know, whatever this weapon is, whatever this dangerous doomsday device is, most likely like a bio weapon, it will be something that is like previously unknown to humans because again, these intelligences will be sort of working at a level that we're unaware of, right?
Like a scientific compound we've not really discovered that's lethal.
Anyway, it will be deadly to 100% of humans and someone will do this.
Someone will, you know, be motivated by this, release this out and not know that they're ending the human race, but they will.
And the idea here is that none of this will take place with any warning, right?
The day that civilization ends will just be the day that it happens because the AI will be crafty enough to hide all this from us.
And the idea here, I think like his primary idea is that this AI doesn't hate us.

It doesn't necessarily want us to die, but it just sees that we are not the most efficient use of Earth, right?
Or the, you know, the energy or the atoms or whatever on Earth.
And it is more intelligent.
And so like, I think, you know, one of the ideas that I heard from somebody else describing this was just basically, it's like, you know, like homo sapiens beat out like other, you know, other competing species, just simply because, you know, we, we were the right ones that we were more intelligent and that that would sort of be what's happening here.
There is a higher intelligence and it is acting, you know, not necessarily maliciously, but it's just acting out of what it thinks is its own rationale.
And that again is the, like that's the, that's the doomsday scenario as, as he sees it or a version of that doomsday scenario.
I just want to like caveat since I've just laid that out in the most science fiction terms.
I don't really see a ton of, I don't really see any like real evidence for that.
It does feel extremely imaginative.
Like it's very, there's not really like a, well, we're right here, you know, in the level of building these things so like it could easily jump to here and then off to the races.
But the thing that I do find just compelling about the argument is AIs writing AIs writing AIs.
Like that is sort of that evolutionary thing where it's like, we don't, you know, you kind of invite chaos if that were to ever happen.
Right.
So, I mean, just to summarize from my own benefit, it's like five steps here.
Step one, design super bacteria, if this technology can, and we hope that it can, be ingenious at coming up with new molecule combinations that can cure diseases, then theoretically the same technology could conceive of molecule combinations that would kill us.
So step one, design super bacteria.
Step two, steal money.
If computer programmers can hack a bank, super ingenious AI can hack an even bigger bank.
Step three, bribe scientists to make this molecular combination.
They got a lot of money.
They can send a plausible email to some science lab and be like, hey, you know, I'm the, whatever, the head of the science department at the, you know, the Charlie and Derek Institute in Germany.
Like, can you please put this together?
I'll pay you, you know, $1.5 million.
Step four, pay a hapless task rabbit to release this bio weapon in wherever.
Step five, everybody dies.
And as you said, it's not about hating humanity.
It is about the logic of select all delete.
Like if you are looking at a page and you want to clear it, you do not care about the letters and the serifs and the spacing and the font.

You're just trying to clear the page, select all delete.
It's that kind of logic that might result in some of these doomsday scenarios.
So let me tell you why I refuse to buy in hook, line and sinker into the Eleizer story.
There's a logic around a lot of AI doomsnares that are kind of like.
Number one, sorry to keep enumerating.
Number one, AI in the future will be able to do anything to within the set of anything includes a lot of bad shit, three, therefore all the bad shit within the set of anything will come to pass.
And it's like, that's actually just a made up syllogism.
It might be predictive, but I don't know how I'd prove it wrong.
You've set up a set of rules for yourself that have no guardrails.
AI will do everything and some things within everything are bad.
I don't know what to do with that, even if that's the best way to think about the future.
To me too, I also just think you mentioned her earlier, this woman I talked to, Melanie Mitchell at the Santa Fe Institute.
And one of the things that she cites in the whole disagreement among experts thing is very simple definitions.
What does it mean for an AI to understand or a large language model to understand?
We don't have definitions around that.
There are some people who believe that understanding is simply the fact that if GPT-4 can ace or get very close to acing the SATs, which are an aptitude test for humans, then it means it has even synthetically an understanding.
The inferences it makes, even though it's not itself conscious, is enough to have replicated human understanding that it understands.
The other side of that group, which is much more like humanist, is saying, no, it does not understand.
And the reason it does not understand is because true understanding is not just making inferences, it's having life experiences.
So it's the experience of being a human in the world and feeling temperature changes.
The example she used with me is if the AI can say the driver angrily cut off the car in front of you or something like that.
But it has no understanding of why someone might have road rage, why that is a dangerous thing to do that.
It doesn't matter that it's making the inference correctly, it doesn't have those experiential qualities.
And those people say then, there's no way that you could get to this level of true artificial general intelligence like a human because it's not going to be able to have those experiences.
The reason I'm saying all this is that I think almost more likely than this idea of truly creating the super being.
I think if you look at the bio lab thing, what's interesting to me there is that can happen just with AI tools as a middleman.
If you're at a lab, a virology lab doing this very controversial gain of function research, and you're saying, hey, we want to push this forward.

If you give this thing the wrong parameters, it could create something horrible.
And I'm talking just about tools that are just trying to come up with some compound and some hapless person engineers based off of a thing because they didn't install the right guardrails in the system that are going to do whatever.
And it creates something and then it gets out in a very organic way and kills a lot of people, maybe doesn't extinguish the human race.
But that's a very easy to imagine problem that doesn't require sentience from the computer.
It just requires a lot of human error based off of a tool that's very powerful.
And I think that's the thing that I am much more likely to be concerned about than Skynet.
When I think about stuff going wrong, I think sometimes the most interesting stories are one big thing going terribly off the rails and ending the human species, but what I'm taking from your point, and I think I agree with it, it's much more plausible to me that many small things go off the rails because of small misalignments or even aligned actors or bad actors with aligned AI create little bad nuisances for us and that generally we perceive over time or maybe even suddenly a weirding of our world.
We come to realize, oh my God, it becomes rote in the news cycle that, oh, well, fucking AI has done something weird again.
There's been another little AI hacking because this regional bank didn't have solid guardrails for its depositors and so there was a hacking and $3.5 million were stolen from First Republic by an AI hacker or there's another corporate scandal, AI hallucinated and Bain told this agriculture AI company that they needed to focus on Chile based on something that was totally made up by the AI.
I think it's much more likely that we see these small crises of misalignment than that we suddenly wake up in a world where half the population has died in the last 12 hours because of some catastrophic misalignment.
A helpful way to think about this is the way that we think about just like a technology as simple as Facebook.
I think if you were to say to somebody in 2007 at Harvard when the Facebook is going around, if someone were to say, this is going to cause a genocide in Myanmar, this will be a primary accelerant for true horrible repression.
People would be like, okay, well, you're out of your mind.
Like truly, we're just trying to figure out the person we met at a party last night.
But I think when you look at how these things happen, it's small, right?
Facebook was not a mind control machine that caused people to go insane and do something.
It was an accelerant for a lot of social conditions based around that left sort of unguardrailed and unchecked helped lead to terrible, awful things in this place and also in a lot of other places across the world.
So I think that's a helpful way to kind of think about these tools, right?
You can say X terrible things going to happen and we open AI might have blood on their hands or something like that, but not necessarily always in the way of, like you said, the big, huge catastrophes.
Sometimes it's simply just a lot of small things going wrong and intersecting with terrible preexisting social, cultural, political currents.

On the more prosaic side, and this is maybe what we can end, I think it's a very interesting question of what kind of jobs can these tools do effectively?
I wonder, do you think chat GPT could do your job?
What parts of a journalist's job do you think chat GPT could do right now?
It's weird, right?
Because I want to be optimistic for myself and a lot of other people and say that it's going to make our jobs weird potentially, but not extinct.
A number of people have made this analogy, but it really does feel to me like the greatest skill that we can all have now is to be editors, right?
Using journalist editor as a thing or in the business world, the way that it's been described to me is chat GPT or GPT-4 or whatever is a really overzealous junior employee, right?
Really smart, really totally capable, no life experience in the field.
It's like give the employee a lot of parameters, let it cook.
It's going to work super hard.
It's going to deliver you something and then you might have to go back and say, okay, well, it's actually like, that's not how things work or whatever it is and edit it and refine it and keep it, manage it.
Something that I watched the GPT-4 open AI demo on I think Wednesday or Tuesday of this week, last week.
The one thing it walked people through was doing taxes and ingesting, you ingest the state's tax code and then you ask some of these really hard questions that you might not know the answer to and it does the calculations, it runs through everything, right?
Then you have to check it.
You have to check its work, right?
You have to go through and make sure and the phrase that they use is like, GPT-4 isn't perfect and neither are you, but together it's going to enhance this thing.
I think that that's a very nice way of putting it for them, right?
It's corporate, I think Microsoft yesterday said that Bing's answers are like usefully wrong, which is my favorite way of putting it, but I do think that there's something there of we're all going to be working together with it.
I think that's right.
I do think that a skill I find being elicited from me when I use GPT-4 is the skill of managing a reflection of my own thoughts.
It puts me in this weird position.
I don't like being a manager.
I used to be an editor for the Atlantic and I kind of self-fired myself from that position because I wasn't very good at it.
I feel like it's activating a muscle that is quite atrophied over the last 10 years, but I'll prompt and it'll give a B plus answer.
I'll say, let's make this A minus together and then sometimes we can scrabble our way towards something that's worthwhile, but it is so weird to think that to sort of round us out here, people, children, adults, young workers, seniors are going to spend the next decade with this little disembodied, super genius daemon next to them, this little assistant

of hallucination and daydreaming and brainstorming and we're going to have to learn the skill of coexisting, of becoming like AI managers in a weird way.
It's like we are all AI managers now.
So Charlie Warzel, thank you very, very much for talking me through this.
I really had fun.
Thanks for having me, man.