I'm Barry Weiss, and this is Honestly.

Six months ago, few people outside of Silicon Valley had even heard of OpenAI, the company that makes the artificial intelligence chatbot known as ChatGPT.

Now, ChatGPT is being used daily by over 100 million users.

And by some of these people, it's being used more often than Google.

Just months after its release, ChatGPT is the fastest growing app in history.

ChatGPT can write essays, it can code, it can ace the barracks in, it can write poems and song lyrics, summarize emails, it can give advice, and it can do all of this in a matter of seconds.

And the most amazing thing of all is that all of the responses it generates are eerily similar to those of a human being.

For many people, it feels like we're on the brink of the biggest thing in human history, that the technology that powers ChatGPT and the emerging AI revolution more broadly will be the most critical and rapid societal transformation in the history of the world.

If that sounds like hyperbole to you, don't take it from me.

What do you compare AI to in the course of human civilization?

You know, I've always thought of AI as the most profound technology humanity is working on, more profound than fire or electricity or anything that we have done in the past.

Google CEO Sundar Pachai said that the impact of AI will be more profound than the invention of fire.

As you know, I work in AI, and AI is changing the world.

Computer scientist and Coursera co-founder Andrew Ying said that AI is the new electricity.

Some compare it to the printing press.

Others say it's more like the invention of the wheel or the airplane.

Saks, you're saying explicitly you think this is bigger than the internet itself, bigger than mobile as a platform shift.

It's definitely top three, and I think it might be the biggest ever.

Many predict that the AI revolution will make the internet seem small.

And last month, the Atlantic ran a story comparing AI to nuclear weapons.

Now, I'm generally an enthusiastic personality, and so when someone tells me about a new technology,

I get excited.

When I heard about crypto, I bought Bitcoin.

When a friend told me that VR is going to change my life, I spent hours trying on his headset in the metaverse.

So there's something profoundly exciting about a technology that so many smart people believe could be a world changer, literally.

We are developing technology which for sure one day will be far more capable than anything we've ever seen before.

But it also scares me because other smart people, sometimes the very same people, are saying that there is a flip side to all of this optimism, and it's a very dark one.

The problem is that we do not get 50 years to try and try again and observe that we were wrong and come up with a different theory and realize that the entire thing is going

to be like way more difficult than we realized at the start.

Because the first time you fail at aligning something much smarter than you are, you die.

One of the pioneers of AI, a guy named Eliezer Yudovsky, claims that if AI continues on its current trajectory, it will destroy life on Earth as we know it.

Here's what he just wrote recently.

If somebody builds a too powerful AI under present conditions, I expect that every single member of the human species and all biological life on Earth dies shortly thereafter.

Now his concerns are particularly severe, it's hard to think of a more dire prediction than that one, but he's not the only one with serious concerns.

Thousands of brilliant technologists, people like Elon Musk and Steve Wozniak, are so concerned that last month they put out a public letter calling for an immediate pause on training any AI systems more powerful than the current version of chat GPT.

So which is it?

Is AI the end of the world, or the dawn of a new one?

To answer that question, I invited Sam Altman on the show today.

Sam is the co-founder and CEO of OpenAI, the company that makes chat GPT, which makes him arguably one of the most powerful people in Silicon Valley, and if you believe the hype about AI, the whole world.

I ask Sam, is the technology that powers chat GPT going to fundamentally transform life on Earth as we know it?

And if so, how?

How will AI affect the way we do our jobs, our understanding of intelligence, our relationships with each other, and our basic humanity?

And are the people in charge of this technology, people like him, ready for the responsibility?

That and more after a short break.

Stay with us.

Hi, honestly listeners, I'm here to tell you about an alternative investing platform called Masterworks.

I know investing in finance can be overwhelming, especially given our economic climate, but there's one thing that will never go in the red, and that is a painting from Picasso's Blue Period.

Masterworks is an exclusive community that invests in blue chip art.

They buy a piece of art, and then they file that work with the SEC.

It's almost like filing for an IPO.

You buy a share representing an investment in the art.

Then Masterworks holds the piece for three to ten years, and then when they sell it, you get a prorated portion of the profit's minus fees.

Masterworks has sold $45 million worth of art to date, from artists like Andy Warhol, Banksy, and Monet.

Over 700,000 investors are using Masterworks to get in on the art market.

So go to masterworks.com slash honestly for priority access.

That's masterworks.com slash honestly.

You can also find important regulation aid disclosures at masterworks.com slash cd.

Sam Altman, welcome to Honestly.

Thanks for having me on.

So Sam, last night I was watching 60 Minutes because despite appearances, I guess I'm a boomer on the inside.

And I listened as Google's CEO compared AI to the invention of fire.

And if that's true, then I guess despite the fact that many of us feel like we're living at the pinnacle of civilization, we're actually in retrospect going to look something like, I guess, Neanderthals or cavemen.

And I wonder if you agree with that analogy, if you think that this technology that you're at the very cutting edge of as the CEO of OpenAI, that it's going to create as seismic a change in human life on Earth as did fire or maybe electricity?

My old understanding of the world did sort of match that there were all of these different technological revolutions and argue about which one is bigger or smaller than the other and talk about when different people reached the pinnacle or whatever.

And now I understand the world in a very different way, which is this one long arc, this one single technological revolution or the knowledge revolution.

And it was our incredible ability to figure things out, to form new explanations in the beginning of infinity language, good explanations and advance the state of knowledge and evolve this sort of infrastructure outside of ourselves, our civilization that really is the way to understand progress forward.

And it's this one big gigantic exponential curve that we're all writing, the knowledge revolution.

And that's now how I view history and certainly how I view the history of technology.

And so I think it's like always tempting to feel like we're at the pinnacle now and I'm sure people in the past felt like they were at the pinnacle and that the part of the revolution that they happen to be living through was the most important part ever.

But I think we are at a new pinnacle and then there will be many more to follow.

And it's all part of this one expanding thing.

But not every period of time feels as enormous.

If you look between like the eighth and 10th century, probably not that much change.

I mean, I'm sure it did to the people who were alive then.

But this feels like a revolution to me in a way that so many other things that have been hyped as revolutions in the past 10 years simply don't.

The curve has squiggles for sure.

And I think this is bigger than some of the things that have been hyped in the last decade that haven't quite planned out.

But that's OK.

That's like the way of it.

Sometimes it looks more obvious.

Sometimes it doesn't.

And again, there are periods where less happens.

But if you zoom out, this is like the third millennium since we've been counting.

But let's say this is maybe year 70,000 of humans or whatever.

And you can say, wow, between years 60 and 70,000 so much happened.
I bet way more will happen between years 70,000 and 80,000 of human history.
I think it's just going to keep going.
Sam, in just a few years, your company has gone from being a small nonprofit that few outside of Silicon Valley paid much attention to, to having an arm of the company that's a multi-billion dollar company with a product so powerful that some people I know tell me they already spend more time on it than they do Google.
Other people are writing off beds warning that the company you're overseeing, the technology overseeing has the potential to destroy humanity as we know it.
For those who are just sort of new to this conversation, what happened at OpenAI over the past few years that's led to what, to many of us seems like this massive explosion only over the past few months.
What have you guys been doing for the past few years?
First of all, we are still a nonprofit.
We have a subsidiary cap profit.
We realized that we just needed way more capital than we could raise as a nonprofit given the compute power that these models needed to be trained.
But the reason that we have that unique structure around safety and sharing of benefits, I think it's only more important now than, than it used to be.
What changed is our seven years, whatever it's been of research, finally really paid off.
It took a long time and a lot of work to figure out how we were going to develop AI and we tried a lot of things.
Many of them came together, some of them turned out to be dead ends.
And finally, we got to a system that was over a bar of utility.
You can argue about whether it's intelligent or not, but most people who use it would not argue that it doesn't have utility.
And then after we developed that technology, we still had to develop a new user interface.
Another thing that I have learned is that making a simple user interface that fits the shape of a new technology is important and usually neglected.
So we had the technology for some time, but it took us a little while to find out how to make it really easy to chat with.
And we were very focused on this idea of like a language interface.
So we wanted to get there.
And then we released that people, it's been very gratifying to see have found a great deal of value in using it to, to learn things, to do their jobs better, to be more creative, whatever.
I know that there are listeners of this show, including my mom, who have vaguely heard what AI is.
They know it's a thing.
They know it's a thing that a lot of people are going on about, either very excited or very scared of it, but they've definitely never used chat GPT.
They've probably never heard of a large language model.

So first, just, just to set the stage, how do you define what artificial intelligence or artificial general intelligence, AGI, is?

What is that?

So I don't like either of those terms, but I've like fought battles in the past to try to change them and given up on that.

So I'll just stick with them for now.

I think AI is understood to still be a computer program, but one that is smarter.

So you still use it like you use some other computer program, but it seems to get what you mean a little bit more.

It seems to be a little bit closer towards like, you know, a smart person that can sort of intuit things or put things together for you in new ways or just be a little bit more natural, a little bit more flexible.

And so people have this experience the first time they talk to chat GPT, which is like, wow, the experts, the linguists, they can argue about definition of the word understanding, but it feels like this thing understands me, feels like this thing is trying to help me and, you know, do my task or whatever.

And that's powerful.

And then AGI is when an AI system gets above a certain threshold.

And we can argue about what that threshold is.

There's a lot of ways to define it.

One that we use sometimes is when it can do more than half of economically valuable human work.

Another one that I really like is when it's capable of figuring out new problems.

It's never seen before when it can sort of come up with brand new things.

A personal one to me is when it can discover new scientific knowledge or more likely help us increase our rate of discovering new scientific knowledge.

But the key difference for someone who hasn't used it is, well, something like Google, right, which changed the world when it came out, trolls the internet for information.

You say to it, you know, Google, find me a story by Sebastian Younger.

You know, what are his best books?

Sorry, I'm looking at a Sebastian Younger book right now.

You know, you could go into GPT and say, write me a story in the voice of Sebastian Younger and seconds later, it can turn that out for you.

Is that right?

Yes.

Yes.

There's a bunch of things that are different, but one that is like totally new is this ability to create.

And again, we can argue about the definition of create and there's many things that can't, but it can give the appearance of creating.

It can put something together for you from things that's already known or seen or understood in a novel way.

And it can leave it to arguing to the computer scientists and the linguists and the philosophers about what it means to create, but from someone getting value out of using

this, which there are a lot of people doing, it does feel like it can generate something new for you.

And this is part of that long arc of technology getting better and better.

Like before you were stuck with whatever a search engine could find in very limited ability to kind of put things together and not always extract information.

And to a user, at least this feels like a significant advancement in that way.

It doesn't even just feel like it.

It is it, right?

Yeah, it is in terms of utility.

There's like a great amount of debate in the field about what are these systems actually doing?

Are we too impressed?

Is it a parlor trick?

But in terms of delivering the value to a user, in some cases, it's inarguably there.

Chat GPT is the fastest growing app ever in the history of the Internet.

In the first five days, it got a million users.

Then over the course of two months after it launched, it amassed a hundred million.

And this was back in January.

And right from the beginning, it was doing amazing things, things that every single dinner party I was going to, it's all anyone could talk about.

It could take an AP test.

It could draft emails.

It could write essays.

I mean, before I went on Bill Maher most recently, I knew we were going to talk about this subject.

I typed in Bill Maher monologue and it churned out a monologue that sounded a whole lot like Bill Maher.

He was not thrilled to hear that.

And yet you have said that you were embarrassed when Chat GPT three and three point five, the first iterations of the product were released.

And I wondered why.

Well, I think the program once said to me, there's always stuck with me is if you don't launch a version one that you're a little embarrassed about, you waited too long to launch.

Explain to Paul Grammys.

Paul Gramm, he ran YC before me and it's just sort of a legend, rightfully so among founders in Silicon Valley.

I think he did more to help founders as a whole, like as a class, probably than any other person, both in terms of starting YC and also just the contributions, the advice and the support he gave to people like me and thousands of other founders he worked with over the years.

Um, but one thing he always said is if you don't launch a version that you're a little embarrassed about, you waited too long.

So there's all of these things in Chat GPT that are still don't work that well.

And, you know, we make it better and better every week.

And that's okay.

Well, last month you released the current version, Chat GPT four, which is remarkably more effective and accurate than the previous versions.

I saw a chart of exam results between Chat GPT three point five and four.

And it's crazy how much better it is.

Like it went from failing the bar exam to getting only 50% of the answers correct to scoring in the 90th percentile or scoring one out of five on an AP calc exam, calculus exam to four out of five, which is much better than I did.

So how were you able at open AI to improve GPT's accuracy with such speed?

And what does that great leap tell us about what the next version of this product will look like?

So we had GPT four done for a long time, but as you said, these technologies are anxiety producing, to say the least.

And when we finished them all, we spent then about eight months aligning it, making it safer, red teaming it, having external audits done.

We really wanted to make sure that that model was safe to release into the world.

And so it felt like it came pretty quickly after three point five, but it was because we had had it for a while and we're just working on safety testing.

Sam, alignment, that word you just used is a word that comes up a lot around this subject. What do you mean when you say it?

That the model acts in accordance with the desire of the person using it and that it follows whatever overall rules have been set for it.

OK, I want to get in a little bit to the safety question, because that's one of the biggest questions people raise.

But just briefly, what are you using this product for right now?

Well, right now, this is the busiest I've ever been in my life.

So right now, I'm mostly using it to help process inbound information.

So summarizing emails, summarizing slack threads, take a very long email.

Someone writes and give me the three bullet point summary.

That kind of stuff I've really come to rely on.

That's probably not its coolest use case, but you asked how I'm personally using it right now, and that's it.

What is its coolest use case?

I'm sure you're hearing from tons of people that are using it.

Give us some examples of the wide range of uses it has.

You know, the one that I find super inspiring, because I just get these heartwarming emails and a lot of them every day are people using it to learn new things and how much it's changed their life there.

You hear this from people in all different areas of the world, all different subjects.

But this idea that with very little effort to learn how to use it this way, you can have a personal tutor for any topic you want and one that really

helps you learn.
That's a super cool thing.
And people really love that.
A lot of programmers rely on it for different parts of their workflow.
Like that's kind of our world.
So we hear about that a lot.
And then we could go on for like a long list of down like every vertical
of what we we've seen there.
There was a Twitter thread recently about someone who says they saved their
dog's life because they input like a blood test and symptoms into GPT for.
That's an amazing use case.
I'm curious where you see chat GPT going.
You know, you use the example of summarizing long-winded emails or
summarizing Slack.
You know, this is kind of like in the, you know, the menial task category, right?
The grocery store order, the sending emails, the making payments.
And then on the other side of it, it's the question about having it do things
that feel more existential and more foundational to what it is to be a human being.
Things that emulate or replace human thinking, right?
So someone recently released an hour long episode of the Joe Rogan experience.
And it wasn't Joe Rogan.
It was someone who created it.
And it was an hour long conversation between you and him.
And the entire thing was generated using AI language models.
So is it the sort of like chores and mindless emails?
Or is it the creation of new conversation, new art, new information?
Because those seem like very different goals with very different human and moral
repercussions.
I think it'll be up to individuals and society as a whole about how we want to
use this technology.
The technology is clearly capable of all of those things.
And it, it's clearly providing value to people in very different ways.
We also don't know perfectly yet how it's going to evolve where we'll hit
roadblocks, what things will be easier than we think, what things will be much,
much harder than we think.
What I hope is that this becomes an integral part of our workflow in many
different things.
So it will help us create.
It will help us do science.
It will help us run companies.
It will help us learn more in school and later on in life.
I think if we change out the word AI for software, which I always like doing,
you know, we say like, is software going to help us create better?

Or is it going to help us do menial tasks better?

Or is that going to help us do science better?

And the answer, of course, is all of those things.

And if we understand AI is just really advanced software, which I think is the right way to do it, then the answer is maybe a little less mysterious.

Sam, in a recent interview, when you were asked about the best and worst case scenarios for AI, you said this of the best case.

I think the best is so unbelievably good that it's hard for me to imagine.

I'd love for you to imagine like, what is the unbelievable good that you believe this technology has the potential to do?

I mean, we can like pick any sort of trope that we want here.

Like what if we're able to cure every disease?

That would be like a huge victory on its own.

What if every person on earth can have a better education than any person on earth gets today? That would be pretty good.

What if like every person, you know, a hundred years from now is a hundred times richer in the subjective sense, better off, like just sort of happier, healthier, more material possessions, more ability to sort of live the good life and the way it's fulfilling to them than people are today?

I think like all of these things are realistically possible.

That was half of the answer that you gave to the question of sort of best and worst case scenarios, right?

I was figuring you're going to mention the other half here.

So here was the other side of it.

You said the worst case scenario was, quote, lights out for all of us.

A lot of people have quoted that line, I'm sure, back to you.

What did you mean by it?

Look, I understand why people would be more comfortable if I would only talk about the great future here.

And I think that's where we're going to get.

I think this can be managed.

And I think the more that we're talking about this now, the more that we're aware of the downsides, the more that we as a society work together on how we want this to go, the way more likely we're going to be in the upside case.

But if we pretend like there is not a pretty serious misuse case here and just say like full steam ahead, it's all great.

Like, don't worry about anything.

I just don't think that's like the right way to get to the good outcome.

You know, as we were developing nuclear technology, we didn't just say like, hey, this is so great, we can power the world.

Like, oh, yeah, don't worry about that bomb thing.

It's never going to happen.

Like the world really grappled with that.

And, you know, it's important that we did.

And I think we've gotten to a surprisingly good place.

There's a lot of people, as you know, who are sort of sounding the alarm bells on what's happening in the world of AI.

Last month, several thousand leading tech figures and AI experts, including Elon Musk, who co-founded OpenAI but left in 2018.

Also, Apple co-founder Steve Wozniak, Andrew Yang, who you backed in the last election.

You're also a UBI fan.

All these people signed this open letter that called for a minimum six month pause on the training of AI systems more powerful than chat GPT-4.

Here's part of what they wrote.

Contemporary AI systems are now becoming human competitive at general tasks.

And we must ask ourselves, should we let machines flood our information channels with propaganda and untruth?

We already have Twitter for that.

Nice.

Should we develop non-human minds that might eventually outnumber outsmart, obsolete, and replace us?

Should we risk they wrote the loss of control of our civilization?

Such decisions must not be delegated to unelected tech leaders.

Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable.

Now, there's two ways, I think, to interpret this letter, at least two ways.

One is that this is a cynical move by people who want to get in on the competition.

And so the smart thing to do is to tell the guy at the head of the pack to pause.

The other cynical way to read it is that by creating fear around this technology, it only makes investments further flood the market.

So I see like two cynical ways to read it.

Then I see a pure version, which is they really think this is dangerous and that it needs to be slowed down.

How did you understand the motivations behind that letter?

Cynical or pure of heart?

You know, I'm not in those people's head, but I always give the benefit of the doubt.

And particularly in this case, I think it is easy to understand where the anxiety is coming from.

I disagree with almost everything about the mechanics of the letter, including the whole idea of trying to govern by open letter.

But I agree with the spirit.

I think we do need, you know, OpenAI is not the company racing right now, but some of the stories we hear from other companies about their efforts to catch up with OpenAI and new companies being started or existing very large ones.

And some of the stories we hear about discussions about being willing to cut corners on safety, I find quite concerning.

What I think we need is a set, and this happens with any new industry, they evolve.

But I think we need an evolving set of safety standards for these models
where before a company starts a training run, before a company releases a new model,
there are evaluations for the safety issues we're concerned about.
There is an external auditing process that happens.
Whatever we agree on as a society are going to be the rules to ensure safe development of this new technology.
Let's get those in place, and you could like pick whatever other technology you want, airplanes.
We have like a robust system for this.
But what's important is that airplanes are safe, not that, you know,
Boeing doesn't develop their next airplane for six months or six years or whatever.
And that's where I'd like to see the energy get redirected.
There were some people who felt the letter didn't go far enough.
Eliezer Yudovsky, one of the founders of the field, or at least he identifies himself that way,
refused to sign the letter because he said it didn't go far enough that it actually understated the case.
I want to read just a few lines to you from an essay that he wrote in the wake of the letter.
Many researchers steeped in these issues, including myself,
expect that the most likely result of building a superhumanly smart AI
under anything remotely like the current circumstances is that literally everyone on Earth will die.
Not as in maybe possibly some remote chance, but as in that is the obvious thing that would happen.
If somebody builds a too powerful AI under present conditions, he writes,
I expect that every single member of the human species and all biological life on Earth die shortly thereafter.
There's no proposed plan for how we would do any such thing and survive.
Open AI's openly declared intention is to make some future AI do our AI alignment homework.
Just hearing that this is the plan ought to be enough to get any sensible person to panic.
And the other leading AI lab he writes, DeepMind, has no plan at all, DeepMind run by Google.
How do you understand that letter?
Someone who doesn't know very much about this subject is reading a brilliant man saying that
every single member of the human species and all biological life on Earth is going to die because of this technology.
Why are some of the smartest minds in tech this hyperbolic about this technology?
Look, I like Eleazar. I'm grateful he exists. He's like a little bit of a prophet of doom.
You know, before AI was going to be nanobots, we're going to kill us all.
And the only way to stop it was to invent AI.
And that's fine, like people are allowed to update their thinking.
I think that actually should be rewarded.
But if you're convinced the world is always about to end and you are not, in my opinion,
close enough to the details of what's happening with the technology, which is very hard in a vacuum,
I think it's hard to know what to do.
So I think Eleazar is super smart.
He may be as smart as you can get about thinking about the problem of AI safety in a vacuum.
The field in general, the field of AI and certainly the field of AI safety has been one of a lot of

surprises.

Things have not gone the way people thought they were going to go.

In fact, a lot of the leading thinkers, I believe including Eleazar, but I'm not sure and it doesn't matter that much.

As recently as like 2016, 2017, we're still not bought into the deep learning approach.

And didn't think that was the thing that was going to work.

And then even if they did, they thought it was going to be like sort of the deep mind RL agents playing games approach.

The direction that things have actually gone or at least are going so far because look, it's almost certainly going to change again,

is that we have these very smart language models that have a lot of properties that, in my opinion, help with the safety problem a lot.

And if you don't consider it that way, if you don't do actual technical hands-on alignment work with the shape of the systems we have

and the risks and benefits that those characteristics lead to, then I think it's super hard to figure out how to solve this problem in a vacuum.

I think this is the case for almost any major scientific or technological program in history.

Things don't work out as cleanly and obviously as the theory would suggest.

You have to confront reality. You have to work with the systems.

You have to work with the shape of the technology or the science, which may not be what you think it should be theoretically, but you deal with reality as it comes.

And then you figure out what to do about that.

A lot of people who are in the AI safety community have said things like,

I never expected that I'd be able to coexist with a system as intelligent as GPT-4.

All of the classical thinking was by the time we got to a system this intelligent,

either we had fully solved the alignment problem or we were totally wiped out, and yet here we are.

So I think the answer is we do need to move with great caution and continue to emphasize, figure out how to build safer and safer systems,

and have an increasing threshold for safety guarantees as these systems become more powerful.

But sitting in a vacuum and talking about the problem in theory has not worked.

Of all of the various sort of doom saying, right, all of the safety or security concerns of these new technologies,

cyber attacks, plagiarism, scams, spreading misinformation, the famous paperclip maximizer thing,

not to mention that this seems like it could be a particularly useful tool for like dictators, warlords, you could think of every scenario.

Which is the one that you, Sam, are most worried about?

I actually find this like a very useful exercise.

So, you know, that quote you just read, like, every person on earth in all biological life is going to totally cease to exist because of AI.

And then I try to like think about how that could happen.

How that would happen. Right. Can you imagine it?

I mean, I could respond if you have some suggestions.

No, like when I read that, I just hear this guy who knows a lot about a technology that I know a

minimal amount about

beyond having used it over the past few months is telling me that it's going to eradicate humanity.

He's not telling me how, but you, I feel like, might have a better understanding of how you could even come to that conclusion.

Well, I don't think it's going to.

I think it is within the, it's within the full distribution in the same way that like nuclear bombs, maybe if we had set all of them off at the same time at the height of the Cold War could eradicate humanity.

But like, I don't think that was most, there were people who made a great name for themselves and a lot of media attention by talking about that.

And I honestly think it's important that they did.

I think having that be such a top of mind issue and having society really grapple with the existential risk of that helped ensure we got to continue to exist.

So I support people talking about it, but I, it's not, it's, again, I think we can manage our way through this fine.

Well, speaking of nuclear, it's been reported that you've compared open AI's ambitions to the ambitions of the Manhattan Project.

I wonder how you grapple with the kind of ethical dilemmas that the people that invented the bomb grappled with.

One of the ones that I think about a lot is the question of, you know, while, while, while the guys that signed that letter calling for the six month pause,

you know, believe that we should pause China, who is using AI already to surveil its citizens, and has said that they want to become the world leader in AI by 2030.

They're not pausing, right?

So make the comparison to me to the Manhattan Projects.

What were the ethical guardrails and dilemmas that they grappled with that you feel are relevant to the advent of AI?

So I think the way that I've made the comparison is that I think the development of AI should be a government project, not a private company project in the spirit of something like the Manhattan Project.

And I really do think that, but given that I don't think our government is going to do a competent job of that anytime soon, it is far better for us to go do that than just like wait for the Chinese government to go do it.

So I think that's what I mean by the comparison, but I also agree with the point you were making, which is we face a lot of very complex issues at the intersection of discovery of new science and geopolitical or deep societal implications that I imagine the team working on the Manhattan Project felt as well.

And so that complexity of like, you know, it feels like we spend as much time debating the issues as we do actually working on the technology.

And I think that's a good thing.

I think it's a great thing.

And I bet it was similar with people working on the Manhattan Project.

Well, right, like in order to ensure that nuclear energy was properly managed after the war, they

created the Atomic Energy Commission, but it took many, many, many people dead.

It took, you know, it took catastrophe in order to set up those guardrails.

Do you think that there will be a similar sort of chain of events when it comes to AI?

Or do you think that we can get to the equivalent of the Atomic Energy Commission before the equivalent of Hiroshima or Nagasaki?

I am very optimistic we can get to it without that happening.

And that's part of the reason that I feel love and appreciation for all of the tumors.

I think having the conversation about the downsides is really important.

Let's talk about the economic impacts of this technology.

You've said that AI systems like GPT will help people live more creatively by freeing up their time, saving them time that they previously used to do boring menial tasks.

But that is going to necessarily result in significant segments of the population, I would imagine, not needing to work.

And the scenario most people imagined was that this technology would first eradicate blue-collar work.

Now it increasingly seems like it will be white-collar work.

It's all the people over here in Hollywood writing television shows.

How do you think it's going to play out? Whose jobs is it going to come for first? Who's second?

And how is it just going to reconfigure the way that we think about work more generally?

Look, I find this issue genuinely confusing.

Even like what we want, I feel, I think we're like confused about whether we want people to work more or work less.

There's a huge debate in France over moving the retirement age two years.

On the other hand, there's a lot of inks spilled by people who have very cushy jobs that get paid a ton about how awful it would be if people who have to work unpleasant minimum wage jobs lose their jobs

were confused on what we even want as the answer here.

We're also confused, as you just pointed out, which is one of my favorite examples of how this is going to impact things.

Experts love to get this wrong. Every pronouncement I have heard about the impact AI is going to have on jobs, it's a question to me of how wrong it sounds.

So I will try to avoid sounding like an idiot a few years in the future and not make a super confident prediction right now.

I'll say the following things.

Number one, the long course of technology is increasing efficiency, often in surprising ways, and thus increasing the leverage of many jobs,

not affecting others as much as you would seem like it would think, and creating new ones that are difficult to imagine before the technology is mature and deployed to the world.

Number two, it seems to me like the human desire to create, to feel useful, to gain status in increasingly silly ways.

That does not seem to me to have any obvious endpoint.

And so the idea that all of us are all of a sudden going to stop working and hang out on the beach all day doesn't feel intuitively right.

But I think the nature of what it means to work and what future society's value will change as it always does.

The jobs of today are very different from the jobs of 200 years ago and very, very different from the jobs of 2,000 years ago.

And that's fine. That's good. That's the way of the world.

The thing that gives me anxiety here is not that we cannot adapt to much better jobs of the future.

We certainly can, but can we do that all inside of one generation, which we haven't had to do in previous technologies?

Sam, you've talked about how AI technologies will, quote, break capitalism.

I've wondered what does that mean and what aspects of capitalism do you think most need to be broken?

Okay, I am super pro-capitalism. I love capitalism. I think it's great.

I do think that over time, the shift of leverage from labor to capital as technology continues gets more and more extreme.

And that's a bad thing. And I can imagine a technology like AI pushing that even further.

And so I believe maybe, not for sure, but maybe we will need to figure out a way to adapt capitalism to acknowledge this fact

that capital has increasing leverage in the world and already has a lot, but it could have much more.

The fundamental precept of capitalism, I think, is still very sound, but I expect it will have to evolve some as it's already been doing.

Let's take a break. How close are we to having AI friends? And is this a technology we should let our kids have? Stay with us.

Okay, let's talk a little bit about the emotional and human concerns that to me are frankly the most interesting.

I talked to Tyler Cowan on the show recently and he thinks that the next generation of kids are going to have essentially what

walking Phoenix has in the movie Her with Scarlett Johansson, like an AI friend or an AI pet or an AI assistant, whatever you want to call that.

And one of the things that parents are going to have to decide is how much time to let their kids spend with their AI

the way our parents had to decide how much TV we were allowed to watch.

I think having a relationship with a bot brings up all kinds of fascinating ethical questions.

The main thing, though, to me is that it's not a real relationship, or maybe you think it is.

No, I don't.

Okay, well, given the amount of time that we see kids already spending on social media, what that's doing for their emotional health,

in what world would having an AI companion be a good step forward for kids?

Oh, I suspect it can easily be a good step forward.

You know, already with what we're hearing about people who are, you know, going through something really hard that they feel uncomfortable talking to their friends about,

or even in some cases they're like uncomfortable or don't have access to a therapist,

and that they're like relying on chat GPT to help them process emotions.

I think that's good.

We'll need some guardrails about how that works, but people are kind of getting very clear and deep value from it.

So I don't know what the guidelines will be.

We'll have to figure out screen time limits or whatever.

But I think there's a role for this for sure.

But don't you fear that given how good the AI is at telling people what they want to hear,

that we can basically create a scenario where everyone is living in their own isolated echo chamber,

and that children aren't developing, especially kids that are born into a world where they're AI natives or whatever,

where they're not learning basic human interactions, basic social skills, how to hear things that they don't want to hear.

To me, it's basically China and kids, or to me the things that when I think about this technology that kind of freaks me out the most.

Yeah, we will need new regulation to prevent companies from following the gradient of hacking attention to get kids to use their product all the time.

But we should address what we're concerned about rather than just say there's no value here when clearly there is.

Okay, one more question about children, and that's the impact that this technology is already having on education.

Some people say that chat GPT has in a matter of months normalized cheating that was already rampant because of COVID,

but normalized cheating among students.

According to this one study I was reading, over a quarter of K through 12 teachers have caught their students cheating with chat GPT,

and roughly a third of these teachers wanted to be banned in their schools.

How much does that worry you, or do you see that as just sort of like we're in the liminal space between the old regime and what we considered fair,

and the new one where this will sort of just be integrated into the way we think about education?

The arc of this has been really interesting to watch, and this both anecdotally matches what I've heard from teachers that I've talked about this,

and also what we've seen from various studies online.

When it initially came out, the reaction was like, oh man, K through 12 education is in a total bad shape.

This is the end of the take home essay, you know, ban it here, ban it there, like it was really like not good.

And now, and it's only been a few months, like five months, something like that.

Now people are very much like, I'm going to change the whole way I teach to take advantage of this, and it's much better than the world before.

And please don't take it away.

A lot of the story of chat GPT getting unbanned in school districts was teachers saying like, this is really important to my kids' education.

And we're seeing amazing things from teachers that are figuring out different ways to get their students to use this or to incorporate this into their classroom.

And, you know, in a world of like very overworked teachers and not enough of them, the fact that there can be supplemental tutoring by an AI system, I think it's really great.

As you definitely know, there has been a lot of discussion over the past few months, heating up, I would say more and more about biases in tech broadly, including at Twitter,

but especially biases in terms of AI, because human beings are creating these programs and therefore the AI is not some like perfect intelligence.

It's built by humans and therefore it's reflecting our biases.

And, you know, the difference somewhat argue between something like Twitter is that we can at least understand the biases and we can follow the people who created the algorithm as they talk back and forth and slack.

But when it comes to a technology like AI, which even its creators don't fully understand how it works, the bias is not as easy to uncover.

It's not as transparent.

How do we know how to find it if we don't know how to look for it?

What do you say to the people who basically look at chat GPT and say, you know, there's bias all over this thing.

And that is unbelievably dangerous.

Forget disinformation.

Forget the creation of propaganda.

The system itself is a kind of propaganda, right?

Elon Musk went on Tucker Carlson.

What's happening is they're training the AI July.

Yes.

It's bad.

To lie.

That's exactly right.

And to withhold information.

To lie and, yes, comment on some things, not comment on other things, but not to say what the data actually demands that I'd say.

How did it get this way?

I thought you funded it at the beginning.

And he claimed that open AI is training the AI as he put it to lie.

What do you make of the conversation around the biases in this technology?

You know, I mentioned earlier that I was embarrassed of the first version of chat GPT.

One of the things I was embarrassed about was I do think the first version did not do an adequate job of representing, say, the median person on Earth.

But the new versions are much better.

And in fact, one thing that I appreciate is most of the loudest critics of the initial version have gone out of their way to say like, wow, open AI listened.

And the new version is much, much better.

We've really looked at our whole training stack to see the different places that bias seeps in biases unavoidable, but find out where it is, how to measure it, how to design evals for it.

Like where we need to give different instructions to human labelers, how we need to get a more

reflective set of human labelers.

And we've made a lot of progress there.

And again, I think it has gone noticed and people have appreciated it.

That said, I really believe that no two people on Earth will ever agree that one AI system is fully unbiased.

And the path here is to set a very broad limits of what the behavior of one of these systems should ever be.

So agree on some things that just we don't do it all.

And that's got to come from society, ideally globally, if it has to be by country, in some cases, which I'm sure it will, that's fine too.

And then B, within that, give each individual user a lot of ability to say, here is the way I want this AI to behave for me.

Here are the things I believe, here's how I would answer this contentious social issue, and the system can then act in accordance with that.

When Elon is saying that open AI is training the AI to lie, is there any truth to that?

I don't even know what he means by that.

You don't have to ask him.

Let's talk a little bit about the ethics of running a company with such potentially world changing technology.

When open AI started, it started as a nonprofit.

And the reason it started as a nonprofit, as you guys articulated it, is that you were concerned about other companies creating potentially dangerous technology purely for profit motivation.

But recently, you've taken that nonprofit and created a capped for profit arm worth $29 billion with a huge investment from Microsoft.

Talk to me about the decision to make that change.

Why did you need to make that change?

That's like how much the computing power for these systems cost.

And we weren't able to raise that as a nonprofit.

We weren't able to raise it from governments.

And that was really it.

I recently read that you have no stake in open AI.

Tell me about the decision to not have any stake in a company that maybe stands to be the most profitable company of all time.

I mean, I already have been super fortunate and done super well.

I have like plenty of money.

This is like the most exciting thing I can imagine working on.

I think it's really important to the world.

This is like how I want to spend my time.

As you pointed out, we started in a way for a particular reason.

And I found that I like personally having like very clear motivations and incentives.

I do think we're going to have to make some very non-traditional decisions as a company.

But I'm like in a very fortunate position of having the luxury of doing this.

Of not having equity.

So you're super rich and so you can make the decision not to do that.

But do you think this technology is so powerful and the incentives,

the possibility of making so much money is so strong that it's sort of an ethical imperative for anyone helming any of these companies

to sort of make the decision to be financially monastic about it?

Like if the incentive in a kind of AI race is to be the first and be the fastest,

you sort of alluded to other companies that are already cutting corners in order to do that, right?

How do you, short of having democratically elected heads of AI companies,

what are the guardrails that can be put in place to prevent people from being corrupted or incentivized in ways that are dangerous?

Actually, I do think democratically elected heads of AI companies or like major AGI efforts, let's say.

I think that is probably a good idea.

Like I don't know why we said short of that.

I think that's like pretty reasonable.

That's probably not going to happen given that the people that are in charge of this country don't even seem to know what substack is.

And you know, like tell me how that would actually work.

I don't know. Like this is all still speculative.

I have been thinking about things in this direction like much more, but like what if all the users of open AI got to like elect the CEO?

It's not perfect, you know, because it impacts people who don't use it and we're still probably too small to have a representative.

We're still way too small to have anything near a representative sample, but like it's better than other things I could imagine.

Okay, well, let's talk a little bit about regulation.

You've said that you can imagine a global governance structure kind of like galactic federation, I guess, that would oversee decisions about the future of AI.

What I would like more than like a global galactic, whatever it is, like something, we talked about this earlier, but something like the IAEA.

You know, something that has real international power by treaty and that gets to inspect the labs, set regulation, make sure we have a cohesive global strategy.

That'd be a great start.

What about the American government right now?

What do you think our government should be doing right now to regulate this technology?

The one thing that I would like to see happen today, because I think it's impossible to screw up and I should just do it, is insight, like government insight, ability to audit whatever training runs models produced above a certain threshold of compute,

or above a certain capability level would be even better.

If we could just start there, then I think the government would begin to learn more about what to do and it would be like a great, a great first step.

I guess my pushback to that would be like, do you really want Diane Feinstein deciding, you know, do you trust the people currently in government even to understand the nature of this technology, let alone regulate it?

I mean, I think you should trust the government more than me, like at least you get to vote them out.
Given that you are the person, though, running it, what are the things that you do to prevent, I guess the word would be like corruption of power that seems to me that it would be the biggest possible risk for you right now?

Like of me personally being corrupted by power?

The company, what do you mean?

Yeah, I mean, well, listen, you've been a very powerful person in your industry for many years.

It seems to me that over the past six months or so, you've become arguably one of the most overseeing a technology that a lot of really smart people are warning at best will completely revolutionize the world,

and at worst will completely swallow it, or as you said, lights out for everybody.

Like how do you deal from, I guess I'm asking a spiritual or an emotional or psychological question.

How do you deal with the burdens of that?

How do you prevent yourself from being, I don't know, like another way of asking that is like what is your North Star?

How do you know that you're making the right choices and decisions?

Well, first of all, I want to like talk about having power.

I don't have, I was going to say I don't have super voting shares, but I don't have shares at all.

I don't have like a special vote.

Like I serve at the pleasure of the board.

I do this the old fashioned way where like the board can just decide to replace the CEO.

I think I'd like to think I would be the first to say if I, for some reason, thought I was not doing a good job.

And I do think, and I don't know what the right way to do this is I don't know what the right timing for it is,

but I do think like whoever is in charge of leading AGI efforts should be democratically elected somehow.

That seems like super reasonable and, you know, difficult to argue with to me.

But it's not like I like have dictatorial power over opening eye nor what I wanted.

I think that's like really important.

That's not what I'm suggesting.

I'm suggesting that like in the firmament of a galaxy that seems like all of the wealth, all of the ideas, all of the,

I don't mean power in the Washington DC sense of it, but power over the future is emanating out of this particular group of people.

And you are one of the stars in that firmament and you've become a brighter and brighter and brighter star.

Like how that's changed you and how you think about...

You mean like the tech industry in general, not like opening eye itself?

I mean tech in general and I mean AI as the sort of pinnacle of the tech world.

It definitely feels surreal.

I heard a story once about, it's always stuck with me for some reason, about this like astronaut, former astronaut that would, you know, years, decades after going to the moon, like stand in his

backyard and look up at the moon
and think it was so beautiful.

And then randomly remember that like, oh fuck, like, you know, decades ago I went up there and walked around on that thing.

That's so crazy.

And I think I sort of hope that's like how I feel about open AI like decades from now.

You know, it's on its 14th democratically elected president or whatever.

And, you know, I'm like living this wonderful life in this fantastic AGI future and thinking about how, you know, marveling at how great it is.

And then I, you know, see something open and I remember that like, oh yeah, I used to run that thing.

But I think like you are probably overstating the degree of power I have in the world as an individual and I probably under perceive it myself.

But, you know, you still just like kind of go about your normal life with all of the normal human drama and wonderful experiences.

And just sort of like the stakes elevate around you or something and you're aware of it and I'm aware of it and I take it like super seriously.

But then like, you know, I'm like running around a field laughing or whatever.

And, you know, you forget it for a little bit and then you remember I'm trying to figure out how to get this across.

It is somehow like very strange and then subjectively not that different.

But I feel the weight of it.

Is there a kitchen cabinet or I guess a signal or WhatsApp group of the people that are in your field talking about these kind of existential questions that this technology is raising?

All the time.

Many signal groups, even across the competitive companies, I think everyone feels the stakes. Everyone feels the weight of it.

Let's talk a little bit about the future and your thoughts on the future.

The computer scientist and futurist Ray Kurzweil predicted in 2017 that AI robots would outsmart human intelligence by 2029.

So I don't know, maybe we'll get there.

He's also been really optimistic about AI's ability to extend our lifespans and heal illness, cure diseases.

He believes by the 2030s we'll be able to augment our brains with AI devices and possibly live forever by uploading a person's neural structure onto a computer or robotic body.

In the Kurzweil vision of the future, where do you fall?

Does that sound realistic to you?

Like it's not prevented by the laws of physics.

So sure, but it feels really difficult to me right now.

You know, we figure everything out eventually, so we'll get there someday, I guess.

There's an idea that has come up a lot over the past while, right?

Just this idea of techno utopianism, this ideology based on the premise that advances in science and technology can sort of bring about something like utopia, right?

By solving depression and cancer and obesity and poverty, even possibly death.

Really, the technology can solve all of our problems.

Do you consider yourself sort of of that school?

Do you believe that technology solves more of our problems than it does create them?

I was going to say, I think technology can solve all problems and continuously create new ones.

So I am a, I'm definitely a pro technologist, but I don't know if I would call myself like a techno utopist.

Is there something that comes to mind that you know technology can't solve?

I do not think that technology can replace genuine human connection in the way I understand it.

One of the things that comes to mind for me when I think about problems that I don't think technology can solve,

but it seems like a lot of smarter people than me disagree is the problem of death.

The average man in the United States born today will live until about 75 years old,

the average woman a little higher, about 80 years old.

You look back to the 1920s, this is an unbelievable improvement.

People then weren't expected basically to live past 55.

You've invested $180 million into a startup called Retro Biosciences,

whose mission is to add 10 years to the human lifespan, putting us at living,

let's call it 85 to nine years old on average.

Tell me why you decided to invest in this and how realistic you think it is

that it's actually going to be able to achieve its goal?

Look, in terms of avoiding biological death, I share your skepticism,

although maybe, you know, if the computer upload whatever, whatever thing works, sure.

More health span, that feels super doable to me.

Like right now, I think our healthcare system, this is part of why I wanted to invest,

is not very good.

We spend a huge amount of money on a low quality of life generally for someone's later years.

And really what you would like, or I think what most people would like,

is to stay very healthy for as long as they can and then have a pretty quick decline,

rather than the way it often happens now.

And that feels to me doable, and I think all of the advances in partial reprogramming

are one of the most exciting things happening in bio right now.

It may turn out to be way harder than we think, it may turn out to be easier,

but it is certainly quite interesting.

For the person who's thinking, what the hell is I'm talking about?

The idea of technology here to extend human life, that just seems so far off.

How can the average person that doesn't have your kind of knowledge and insight into technology

prepare for what is about to come over the next five or 10 years?

Before this interview, I went on Twitter and I asked people what I should ask you.

And there was a Twitter user, Alex, who wrote, if you were a college senior,

what majors and career pathways would you decide or would you recommend, Sam,

knowing what's around the bend, in light of AI development especially?

I think it's a big mistake to put too much weight on advice from other people.

In my life, I have been steered badly by advice much more often than the other way around.

So you don't give advice ever?

I think I used to give too much advice because it was such a part of running YC or being a YC partner.

And now I try to give much less advice with much more awareness of how frequently advice is wrong.

Study whatever you want.

Like follow your own personal curiosity and excitement,

realizing the rate of change in the world is going to be high

and that you need to be very resilient to such change,

but don't take your life advice about what to go work on from somebody else.

There have been a lot of moments in the past decade where people said

a new technology was going to completely upend the world as we know it.

They said that about virtual reality. They said it about crypto.

And personally, I don't own a VR headset and I have $10,000 in Bitcoin

that I don't know how to get out because I forgot my Coinbase password.

I think the question a lot of people are wondering is what makes this different?

Well, we might be wrong, right?

Like they might be right. This might not be different.

This could hit a wall. This could change things somehow much less than we think.

Even if AI is really powerful, it might just mean the world goes much faster,

but the human experience doesn't change that much.

I'm very biased.

My personal belief for the last decade has been that the two most important technological trends would be AI and abundant energy.

And I've spent all my time on those things.

And it's very much what I believe in.

And it's very much like my filter bubble.

So I think that's right.

But I think anyone listening should have a huge amount of skepticism on me saying that.

And it might not be different.

I mean, hopefully it's going to be better than crypto and the metaverse.

But even those, I think are going to be pretty cool.

Another project that I work on is this thing called Worldcoin

that I helped put together a few years ago.

It was like horribly mocked for a long time.

And now all of the kind of like crypto tourists have gone.

The true believers are still there.

People see why we wanted to start the project.

And now it's like, I think, super exciting.

So, you know, it's just like the future is hard to predict.

These trends take a while to untangle.

Sam Altman, let's do a lightning round.

All right.

Sam, what is the best thing you've ever invested in?

I mean, the most joy.

Joy. Let's go joy.

All of the time spent in open AI.

Okay. And financially?

I suspect that'll turn out to be helium.

What is helium?

It's a nuclear fusion company.

I'm pretty closely involved with.

What is the first thing you ever asked chat GPT?

That is a good question.

I don't remember.

I think it most likely would have been some sort of arithmetic question.

Sam, do you think UFOs are real?

Like, do I think they're aliens or do I think there's been like flying objects

from other militaries that we don't know what they are?

Not flying objects. Do you think that there are aliens?

No.

What do you look for when you're interviewing for a candidate

applying for a job at open AI?

All of the normal things that I would look for for any other role,

you know, intelligence, drive, hard work, creativity, team spirit,

all of the normal things, plus a real focus and dedication

to the good AI outcome.

What is one book that you think everybody should read?

I mentioned it earlier in this conversation about, I'll say, the beginning of infinity.

I know you don't like advice, but what's the best piece of advice

that you've ever received?

Don't listen to advice too much.

What is a fundamental truth that you live by?

You can get more done than you think.

You are like capable of more than you think.

You get to have dinner tonight with anybody dead or alive.

Your dream dinner. Who's at that dinner?

I think I'd have a very different answer to this question,

like any day given like what I'm thinking about,

but you like what my pick for today.

Yeah, today.

Today I pick Alan Turing.

Interesting.

A few years ago, you told a colleague, and it was in the New Yorker,

great profile about you, that you were ready for the end of the world.

You sort of outed yourself as a prepper.

You had gold, you had batteries, you had a patch of land in Big Sur.

Are you still a prepper?

No, not in the way I would like think about it.

It was like a fun hobby, but there's nothing else to do.

I also like for all of this stuff about like, oh man, like, you know,

none of this is going to help you if AGI goes wrong.

But it's like a fun hobby.

Sam, you grew up Jewish. Do you believe in God?

Uh, I want to say yes, but not in the Jewish God

or the way that I think most other people would define that question.

What do you mean by that?

Um, I can't answer this in a lightning round.

Okay, here's some questions from chat GPT.

Sam, GPT wants me to ask,

what futuristic technology do you wish existed today?

Can I say AGI?

Sure.

What technology do you think will be obsolete in 10 years?

GPT-4.

What futuristic mode of transportation are you most excited about?

Fusion-powered starships.

And Sam, last question brought to you by your own company.

When were you first introduced to AI?

And what about the concept stuck with you?

What made you believe in its potential?

I must have heard about it first from sci-fi,

but my subjective memory of this is as a child using a computer

thinking about what would happen when the computer could think.

How old were you?

Eight.

There's a million more questions I want to ask you,

but we're out of time and I know you need to go

and do a lot of things at OpenAI.

So Sam Altman, thanks for joining us.

Thanks for having me on.

Thanks for listening.

We think AI is an unbelievably interesting topic,

one we want to cover more on the show.

If you were provoked by this conversation,

if it educated you, if it excited you, if it concerned you,

if it made you want to go and use chat GPT

and find out what it's all about, it's great.

Share this conversation with your community
and use it to have a conversation of your own.
And if you want to support honestly,
there's just one way to do it.
Subscribe by going to thefp.com today.
See you next time.