Artificial intelligence is superhuman.

It is smarter than you are, and there's something inherently dangerous for the dumber party in that relationship.

You just can't put the genie back in the bottle.

Damn Harris!

Neuroscientist, philosopher, author, podcaster.

He goes into intellectual territory where few others dare tread.

Six years ago, you dared talk.

The gains we make in artificial intelligence could ultimately destroy us.

If your objective was to make humanity happy, and there was a button placed in front of you, and it would end artificial intelligence, what would you do?

Well, I would definitely pause it.

The idea that we've lost the moment to decide whether to hook our most powerful AI to everything is just,

oh, it's already connected to the internet, got millions of people using it.

And the idea that these things will stay aligned with us because we have built them,

yet we gave them a capacity to rewrite their code.

There's just no reason to believe that.

And I worry about the near-term problem of what humans do with increasingly powerful AI,

how it amplifies misinformation.

Most of what's online could soon be fake.

Can we hold a presidential election 18 months from now that we recognize as valid?

Like, is it safe?

And it just gets scarier and scarier.

I worry we're just going to have to declare bankruptcy to the internet.

The internet, the internet.

If your intuition is correct, are you optimistic about our chances of survival?

Quick one before this episode starts.

Two years ago, I started writing a book based on everything I've learned from doing this podcast and meeting all of these incredible people, and also from my career in business.

from running my marketing agency Flight Story, my software business ThirdWeb,

my investment fund called Flight Fund and everything else I've been doing in business and in life.

From this, I've created a brand new book called The Diary of a CEO.

33 Laws for Business and Life.

If you want to build something great or become great yourself like the guests I get to interview, I ask you, please, please, please read these 33 laws.

The book I always should have written.

If you like this podcast, this book is for you.

The link is in the description below, wherever you're listening to this.

You can get the pre-order right now on Apple Books and it'll be delivered this month.

If you want to listen to the book, you can pre-order the audio right now on Apple Books.

Sam, six years ago you did a TED Talk.

I watched that TED Talk a few times over the last week and the TED Talk was called

Can We Build AI Without Losing Control Over It?

In that TED Talk, you really discuss the idea whether AI, when it gets to a certain point of sentience and intelligence, will wreak havoc on humanity.

Six years later, where do you stand on it today?

Do you think, are you optimistic about our chances of survival?

Yeah, I can't say I'm optimistic.

I am worried about two species of problem here that are related.

There's the near-term problem of just what humans do with increasingly powerful AI and how it amplifies the problem of misinformation and disinformation and just makes it harder to make sense of reality together.

Then there's just the longer-term concern about what's called alignment with artificial general intelligence, where we build AI that is truly general and by definition superhuman in its competence and power.

Then the question is, have we built it in such a way that is aligned in a durable way with our interests?

There are some people who just don't see this problem.

They're kind of blind to it.

In the presence of someone who doesn't share this intuition, they don't resonate to it,

I just don't understand what they're doing or not doing with their minds in that moment.

Let's say I'm wrong about that.

It's just the other person's right.

We just have fundamentally different intuitions about this particular point.

The point is this, if you're imagining building true artificial general intelligence that

is superhuman and that is what everyone, whatever their intuitions, purports to be imagining here.

There are people on both sides of the alignment debate.

There are people who think alignment is a real problem and people who think it's a total fiction.

Firstly, everyone who's party to this conversation agrees that we will ultimately build artificial general intelligence that will be superhuman in its capacities.

There's very little you have to assume to be confident that we're going to do that.

There's really just two assumptions.

One is that intelligence is substrate independent.

It doesn't have to be made of meat.

It can be made in silico.

We've already proven that with narrow AI.

We obviously have intelligent machines.

Your calculator in your phone is better than you are at arithmetic and that's some very narrow band of intelligence.

As we keep building intelligent machines on the assumption that there's nothing magical about having a computer made of meat,

the only other thing you have to assume is that we will keep doing this.

We will keep making progress and eventually we will be in the presence of something more intelligent than we are.

That's not assuming Moore's law.

It's not assuming exponential progress.

We just have to keep going.

When you look at the reasons why we wouldn't keep going, those are all just terrifying because intelligence is so valuable

and we're so incentivized to have more of it and every increment of it is valuable.

It's not like it only gets valuable when you double it or 10X it.

No, no, if you just get three more percent, that pays for itself.

We're going to keep doing this.

Our failure to do it suggests that something terrible has happened in the meantime.

We've had a world war.

We've had a global pandemic far worse than COVID.

We got hit by an asteroid.

Something happened that prevented us as a species from continuing to make progress in building intelligent machines.

So absent that, we're going to keep going.

We will eventually be in the presence of something smarter than we are.

This is where intuitions divide.

My intuition, and it's shared by many people, I know at least one who you've spoken to, my intuition is that there is something inherently dangerous for the dumber party in that relationship.

There's something inherently dangerous for the dumber species to be in the presence of the smarter species.

We have seen this based on our entanglement with all other species, dumber than we are, or certainly less competent than we are.

Reasoning by analogy, it would be true of something smarter than we are.

People imagine that because we have built these machines, that is no longer true.

Here's where my intuition goes from there.

That imagination is born of not taking intelligence seriously.

Because what intelligence is, is a mismatch in intelligence in particular,

is a fundamental lack of insight into what the smarter party is doing and why it's doing it, and what it will do next on the part of the dumber party.

You just could imagine that by analogy, just imagine that dogs had invented us as their superintelligent AIs.

For the purpose of making their lives better, just securing resources for them,

securing comfort for them, getting them medical attention.

It's been working out pretty well for the dogs for about 10,000 years.

There's some exceptions. We mistreat certain dogs.

Generally speaking, for most dogs, most of the time, humans have been a great invention.

It's true that the mismatch in our intelligence dictates a fundamental blindness with respect to what we've become in the meantime.

We have all these instrumental goals and things we care about that they cannot possibly conceive.

They know that when we go get the leash and say, it's time for a walk, they understand that particular part of the language game.

But everything else we do when we're talking to each other, when we're on our computers or on our phones,

they don't have the dimmest idea of what we're up to.

The truth is, we love our dogs. We make irrational sacrifices for our dogs.

We prioritize their health over all kinds of things that it's just amazing to consider.

Yet, if there was a new global pandemic kicking off and some Xenovirus was jumping from dogs to humans,

and it was just super Ebola, it was 90% lethal, and this was just a forced choice between them.

What do you value more? The lives of your dogs or the lives of your kids?

If that's a situation we were in, it's totally conceivable, by no means impossible.

We would just kill all the dogs, and they would never know why.

It's because we have this layer of mind and culture and just the new sphere.

There's this realm of mind that requires a requisite level of intelligence to even be partied to, to even know exists.

that they have no idea it exists.

This is a fanciful analogy because the dogs did not invent us, but evolution invented us.

Evolution has coded us, as I said, to survive and spawn, and that's it.

Evolution can't see everything else we've done with our time and attention and all the values we've formed in the meantime.

And all the ways in which we have explicitly disavowed the program we've been given.

Evolution gave us a program, but if we were really going to live by the lights of that program, what would we be doing?

We would be having as many kids as possible.

Guys would be going to sperm banks and donating their sperm and finding that the best use of their time and attention.

The idea that you could have hundreds of kids for which you have no financial responsibility, that should be the most rewarding thing that you could possibly do with your time as a man.

And yet, that's obviously not what we do, and there are people who decide not to have kids.

And everything else we do, from having podcast conversations like this to curing diseases, literally everything we're doing with science, with culture, is, yes, there are points of contact between those products and our evolved capacities.

It's not magic. We are social primates that have leveraged certain ancient hardware to do new things, but the code that we've been given doesn't see any of that.

And we've not been optimized to build democracies.

Evolution knows nothing. It can know nothing. If evolution were a coder, there's just no democracy maximization in that code. It's just not there.

So the idea that these things will stay aligned with us because we have built them, because if we have this origin story that we gave them their initial code, and yet we gave them a capacity to rewrite their code and build future generations of themselves, right?

There's just no reason to believe that. And the mismatch in intelligence is intrinsically dangerous. And you could see this by, I mean, Stuart Russell, I don't know if you had him on the podcast. He's a great professor of computer science at Berkeley, and he wrote, literally co-wrote one of the most popular textbooks on AI.

He has some arresting analogies, which I think are good intuition pumps here.

And one is just think of how you would feel if you knew, like, let's say we got a communication from elsewhere in the galaxy, and it was a message that we decoded and it said, people of Earth, we will arrive on your lowly planet in 50 years.

Get ready, right?

Anyone who thinks that we're going to get super intelligent AI in, let's say, 50 years thinks we're essentially in that situation, and yet we're not responding emotionally to it in the same way.

If we received a communication from a species that we knew just by the sheer fact that they were communicating with us in this way, we knew they're more competent and more powerful and more intelligent than we are, right?

And they're going to arrive, right? We would feel that we were on the threshold of the most momentous change in the history of our species.

And we would feel, but most importantly, we would feel that it's because this is a relationship, an unavoidable relationship that's being foisted upon us, right?

It's like a new creature is coming into the room, right, with its own capacities, and now you're in relationship.

And one thing is absolutely certain, it is smarter than you are, right?

By what factor? I mean, ultimately, we're talking about by factors, you know, just by so many orders of magnitude, our intuitions completely fail.

I mean, even if it was just a difference in the time of processing, let's say there was no difference in the actual native intelligence, but it's just processing speed.

A million-fold difference in processing speed is just a phantasmagorical difference in capacity. Just imagine we had 10 smart guys in a room over there, and they were working and thinking and talking a million times faster than we are, right?

Well, so they're no smarter than we are, but they're just faster, and we talk to them once every two weeks just to catch up on what they're up to and what they want to do and whether they still want to collaborate with us.

Well, two weeks for us is 20,000 years of analogous progress for them.

So how could we possibly hope to constrain the opinions and collaborate with and negotiate with people no smarter than ourselves who are making 20,000 years of progress every time we make two weeks of progress, right?

It's unimaginable, and yet there are many people who just think this is just fiction.

All the noises I've made in the last five minutes are just like a new religion of fear, right?

And it's just there's no reason to think that alignment is even a potential problem.

If your intuition is correct, and that analogy of us getting a signal from outer space that someone is coming in 30 years, which by the way, a lot of people that speak on this subject matter don't believe it's even going to be 30 years until we reach that sort of singularity moment.

I think they speak of artificial general intelligence.

I've heard people like Elon say, you know, many fewer decades, 10 years, 15 years, 20 years, etc.

If that is correct, then surely this is the most pressing challenge, conversation issue of our time.

And there's no logical reason that I can see to refute your intuition there.

I can't see a logical reason.

The rate of progress will continue.

Don't necessarily see anything that will wipe out or pause our rate of progress.

Let me just be charitable to the other side here.

There are other assumptions that they smuggle in that some do it without being aware of it, but some actually believe these assumptions.

And this spells the difference on this particular intuition.

So it's possible to assume that the more intelligent you get, the more ethical you become by definition right now.

And we might draw a somewhat more equivocal picture from just the human case where we see that, oh, there's some very smart people who aren't that ethical.

But I believe there are people, and I've talked to at least a few people who believe this.

There are people who assume that kind of in the limit as you push out into just far beyond human levels of intelligence,

there's every reason to believe that all of the provincial, creaturely failures of human ethics will be left behind as well.

It's like you're not like the selfishness and the basis for conflict.

These are not the apish urges of status seeking monkeys is just not going to be in the code.

And as you push out into just kind of the omnibus genius of the coming AI, there's a kind of a sainthood that's going to come along with it and a wisdom that will come along with it.

Now, I just think that's quite a gamble.

I would take the other side of that bet and I was framed it this way.

There have to be ways in the space of all possible intelligences that are beyond the human.

There's got to be more than one possible.

It's just like there's many different ways to have a chess engine that's better than I am at chess.

They're different from each other, but they're all better than me, right?

There's got to be more than one way to have a superhuman artificial intelligence.

And I would imagine there are, you know, not an infinite number of ways, but just a vast number of... In the space of all possible minds, there are many locations in that space beyond the human that are not aligned with human wellbeing, right?

There's got to be more ways to build this unaligned than aligned, right?

And what other people are smuggling into this conversation is the intuition that, no, no, once you get beyond the human, it's just going to get...

You're going to be in the presence of, you know, just the Buddha who understands quantum mechanics and oncology and everything else, right?

I just see no reason to think that that's so, and we could build something that is, again, taken intelligence seriously.

We're going to build something that we're in relationship to.

It's really intelligent in all the ways that we're intelligent.

It's just better at all of those things than we are.

It's, by definition, superhuman because the only way it wouldn't be superhuman, the only way it would be human level, even for 15 minutes,

is if we didn't let it improve itself, if we wanted to just keep it stuck at, you know, at a...

We built a college undergraduate and we wanted just to keep it stuck there,

but we would have to dumb down all of the specific capacities we've already built, right?

Every AI we have, narrow AI, is superhuman for the thing it does.

It has access to all the information on the Internet, right?

It's just like it's got perfect memory.

It can perfectly copy itself.

When one part of the system learns something, the rest of the system learns it because it just can swap files, right?

It's, again, your phone is a superhuman calculator.

There's no reason to make it a calculator that is human level.

And so we're never going to do that.

We're never going to be in the presence of human AGI.

We will be immediately in the presence of superhuman AGI.

And then the question is how quickly it improves and how farther, how much headroom is there to improve into.

On the assumption that you can get guite a bit more intelligent than we are, right,

that we're nowhere near the summit of possible intelligence,

you have to imagine that you're going to be in the presence of something that is, again, it could be completely unconscious, right?

I'm not saying that there's something that's like to be this thing, although there might be.

And that's a totally different problem that's worth worrying about.

But whether conscious or not, it is solving problems, detecting problems,

improving its capacity to do all of that in ways that we can't possibly understand.

And the products of its increasing competence are always being surfaced, right?

So it's like we've been using it to change the world.

We've become reliant upon it.

We built this thing for a reason.

I mean, one thing that's been amazing about developments in recent months is that those of us who have been at all cognizant of the AI safety space for now going on a decade or more, for some people,

always assumed that as we got closer to the end zone, that the labs would become more circumspect.

We'd be building this stuff air-gapped from the internet.

It's like we have this phrase air-gapped from the internet.

We thought this was a thing.

This thing would be in a box.

And then the question would be, well, do we let it out of the box and let it do something?

Like, is it safe?

And how do we know if it's safe?

And we thought we would have that moment.

We thought it would happen in a lab at Google or at Facebook or somewhere.

We thought we would hear, OK, we've got something really impressive,

and now we just want it to touch the stock market, or we want it to touch our medical data, or we just want to see if we can use it.

We're way past that.

We've built this stuff already in the wild.

It's already connected to the internet.

It's already got millions of people using it.

It already has APIs.

It's already doing work.

So from an AI safety point of view, that's amazing.

We didn't even have the choice point we thought was going to be so fraught.

Of course we didn't.

Because there was such pressing incentives for people to press forward regardless of that conversation,

especially you.

But everyone thought, I mean, I don't believe I was ever in conversation with someone.

I mean, someone like L.A.'s or Yudikowski or Nick Bostrom or Stuart Russell,

who assumed we would be in this spot.

Because, again, I'd have to go back and look at those conversations.

But there was so much time spent.

It seems quite unnecessarily on this idea that we'd make a certain amount of progress, and circumspection would kick in.

Even the people who were doubters would become worried.

And there would be, like in the final yards, as we go across into the end zone,

there'd be some mode where we could sort of slow down and figure it out,

like try to deal with the arms race dynamics.

Place a phone call to China and just like, let's talk about this.

We've got something interesting.

But the stuff is already being built in connection to everything.

And there's already just endless businesses being devised on the back of this thing.

And all the improvements are going to get plowed into it.

So just imagine what this looks like even in success.

Like, let's say it just starts working wonders for us,

and we get these great productivity gains.

OK, so then we cross into whatever the singularity is at whatever speed

we find ourselves in the presence of something that is truly general.

After all of this stuff, all of this narrow stuff,

albeit superhuman, narrow stuff, is something that we totally depend on.

Like every hospital requires it and every airplane requires it

and all of our missile systems require it.

And this is the way we do business.

There is nothing to turn off at that point.

I put this to Mark Andreessen on my podcast and he said.

yeah, you can turn off the internet.

I can't believe he was quite serious.

I mean, yes, if you're North Korea, I guess you can turn off the internet for North Korea.

And that's why North Korea is like North Korea.

But the idea that we could just the cost of turning off the internet now would be,

I think it would be unimaginable in the economic, just the economic cost alone.

So anyway, I mean, just the idea that we've lost the moment to decide

whether to hook our most powerful AI to everything,

because it's already being built more or less in contact with, if not everything,

so many things that you just can't put the genie back in the bottle.

That is genuinely surprising to me.

And yeah, I mean, incentives to tell the tale.

Is this not the most pressing problem, though?

Because I was going to ask you this conversation by asking you the question

about the thing that occupies your mind the most

and the most important thing we should be talking about.

And I in part assumed the answer would be artificial intelligence,

because the way that you talk about your intuition on this subject matter, you've got children.

You think about the future a lot.

If you can see this species coming to Earth in the next,

even if it's in the next 100 years,

it strikes me to be the most pressing problem for humanity.

Well, I do, as interesting as I think that problem is and consequential as it is,

I'm worried that life could become unlivable in the near term before we even get there.

Like, I'm just worried about the misuses of narrow AI in the meantime.

I'm worried about, just take the current level of AI we have.

We have GPT-4.

I think within the next 12 months or two years,

let's say whatever GPT-5 is,

we're going to be in the presence of something where most of what's online

that purports to be information could soon be fake.

Most of the text you find on any topic is just fake.

Someone has just decided,

write me a thousand journal articles on why mRNA vaccines cause cancer

and give me 150 citations, write them in the style of nature

and nature genetics and Lancet and JAMA and just put them out there.

One teenager could do that in five minutes with the right AI.

GPT-4 is not quite that, but GPT-5 possibly will be that.

That is such a near-term advance.

When you imagine knitting together the visual stuff like mid-journey and dolly

and stable diffusion with a large language model,

just imagine the tool.

Again, maybe this is 18 months away.

Maybe it's three years away, but it's not 30 years away.

The tool where you can just say,

give me a 45-minute documentary on how the Holocaust never happened.

Filled with archival imagery,

give me Hitler speaking in German with the appropriate translations and give it in the style of Alex Gibney or Ken Burns and give me 10,000 of those.

All the friction for misinformation has been taken out of the system.

I worry we're just going to have to declare bankruptcy with respect to the internet.

We just are not going to be able to figure out what's real.

When you look at how hard that is now with social media in the aftermath of COVID and Trump and just the challenge of holding an election that most of the population agrees was valid.

That challenge already is on the verge of being insurmountable in the US.

It's easy to see us failing at that, AI aside.

When you add large language models to that and the more competent future version of it, where it's just the most compelling deep fakes are indistinguishable from real data.

Everyone is siloed into their tribes where they're stigmatizing the information that comes from any other tribe.

The internet is now so big a place that there really isn't the ordinary selection pressures where bad information gets successfully debunked so that it goes away.

You can live in a conspiracy cult for the rest of your life if you want to.

You can be queuing on all day long if you want to.

Now we've got deep fakes shoring all that up and just spurious scientific articles shoring all that up. All of this just becomes a more compelling form of psychosis and culturally speaking.

I'm just worried that it's going to get harder and harder for us to cooperate with one another and collaborate

and that our politics will just completely break and that'll offer an opportunity for lots of bad actors. There's cyber-terrorism and there's synthetic biology that the moment you turn AI loose on the prospect of engineering viruses

the asymmetry here is that it seems like it's always easier to break things than to fix them or to categorically prevent people from breaking them.

What we have with increasingly powerful technology is the ability for one person to create more and more damage

or one small group of people.

It just turns out it's hard enough to build a nuclear bomb that one person can't really do it no matter how smart.

You need a team and traditionally you need state actors and you need access to resources and you need to have to get the fissile material and it's hard enough.

But this is being fully democratized, this tech.

I worry about the near-term chaos.

I've never found the narrow-term consequences of artificial intelligence to be that interesting until now.

That image of the internet becoming unusable, so that was a real eureka moment for me because I've not been thinking about that.

Yeah, me too. I was just concerned about the AGI risk and now really in the aftermath of Trump and COVID

I see the risk of, you know, if not losing everything, losing a lot that matters just based on our interacting with these very simple tools that are reliably misleading us.

I'm amazed at what social media, I forget about, I'm amazed at what Twitter did to me.

Even with all of my training and with my head screwed on reasonably straight.

It's amazing to say it, but almost all of the truly bad things that have happened to me in the last decade

that just really like just destabilized relationships and just priorities and really kind of got plowed back into me.

It became a kind of professional emergency, you know, stuff I had to respond to, you know, in writing or on podcasts.

It was all Twitter, my engagement with Twitter was the thing that produced the chaos and it was completely unnecessary and it was just, it was amplifying a kind of signal for me that I felt compelled to pay attention to because I was on it

and I was trying to communicate with people on it and I was getting certain communication back and it was giving me a picture of the rest of humanity, which I now think is fundamentally misleading

but it was still consequential in its, even believing, at a certain point believing that it was misleading wasn't enough to inoculate me against the delusion of the opinion change that was being forced upon me

and I was feeling like, okay, these people are becoming unrecognizable.

I know some of these people, I've had dinner with some of these people

and their behavior on Twitter is appearing so deranged to me

and so in such bad faith, the people who I know to be non-psychopaths

are starting to behave like psychopaths, at least on Twitter

and I'm becoming similarly unrecognizable to them.

It's just, again, it all felt like a psychological experiment to which I hadn't consented and which I enrolled myself somehow because it was what everyone was doing in 2009 and I spent 12 years there getting some signal and responding to it and it's not to say that it was all bad.

I read a bunch of good articles that got linked there and I discovered some interesting people but the change in my life after I deleted my Twitter account was so enormous.

I mean, it's embarrassing to admit it.

It's like getting out of a bad relationship.

It was a fundamental freedom from this chaos monster that was always there ready to disrupt something based on its own dynamics.

When did you delete it?

I think it was December.

I'm not someone that really takes sides on things.

I like to try and remain in the middle.

So you must have a very different Twitter experience than I was having? No.

So I don't tweet anything other than this podcast trailer.

I don't tweet anything else.

The only thing you'll see on my Twitter is the podcast trailer.

That's it.

And for all the reasons you've described and more interestingly,

I wanted to say in the last eight months as someone that doesn't get caught up too much in the media

or Elon bought this, it's 100% gone in that direction.

As in my timeline now is, I say it to my friends all the time

and some of my friends who are again, I think are nuanced and balanced have said to me,

there's something that's been turned up in the algorithm to increase engagement

that has planted me in an unpleasant echo chamber that I didn't desire to be in.

And if I wasn't somewhat conscious, I would 100% be in there.

My timeline, my friend tweeted the other day, my friend Cahill tweeted,

he's never seen more people die on his Twitter timeline than he has in the last six months.

They're prioritizing video.

So you're seeing a lot of like death and CCTV footage that I've never seen before.

And then the debate around gender, politics, right-leaning subject matter

has never been more right down your throat.

Because it's almost like something in the algorithm has been switched

where it's now like people have been let out of the asylum.

That's the only way I can describe it.

And it's made me retract even more.

So when Zuckerberg announced threads the other couple of weeks ago,

it was kind of like a life raft out of the Titanic.

And I really, really mean that.

And I'm not someone to get easily caught up in narrative, you know,

as it relates to social media platforms, it's been my industry for a decade.

But what I've seen on Twitter, and it's actually made me believe this hypothesis

I had five years ago where I thought they would be,

I thought the journey of social networking would have way more social networks and they'd be more siloed.

I thought we'd have one for our neighborhood, our football club.

And now I believe that even more than ever.

Yeah, that seems right.

And I think whether it's possible to have a truly healthy social network

that people want to be in and it's a good reason to be there.

And it's, I don't know if it's possible.

I like to think it is, but it's, I think there are certain things

you have to clean up at the outset that is supposed to make it possible.

And I think anonymity is a bad thing.

I think probably being free is a bad thing.

I think, you know, you sort of get what you pay for online.

And if it's, I just think there might be ways to set it up that way would be better.

But I didn't think it'd be popular.

I think with the thing that makes it popular makes it toxic.

Right.

Right.

And even the anonymity piece, I've played this out a couple of times in my mind and the rebuttal I always get is while there's people in Syria

who have news to break important news to break and they'd be hung if they.

So we need an anonymous version of the social internet.

Right.

Yeah.

Well, I guess there could be some exception there, but I don't know.

I don't know.

It just doesn't, it actually doesn't interest me because I just feel such a different sense of my being in the world as a result of not paying attention to the,

my online, the simulacrum of myself.

It's a, because Twitter was the only one I used.

Like I was on, I've been on Facebook this whole time.

I've been on, I think, I guess I'm on Instagram too.

But it's like my team just uses those as marketing channels.

You know, it's just like, it sounds like that's the way you use Twitter now.

But Twitter was the one that I decided, okay, this is going to be me.

I'm going to be posting here.

I'm going to, you know, if, if I've made a mistake, I want to hear about it.

You know, it's like, and I just wanted to use it as, as actual,

actual basis for communication.

And for the longest time, it actually felt like a valid tool in that respect.

You know, it reached a crisis point.

I decided this is just pure toxicity.

There's just no reason, even the good stuff can't possibly make a dent in the bad stuff.

So I just deleted it.

And then I was, I was returned to the real world, right?

Where I've, where I actually live and to books and to, I mean, I'm online all the time anyway,

but, but it's not having the, it's the time course of reactivity.

When you don't have social media, when you don't, and you don't have a place to put this,

this instantaneous hot take that you're tempted to put out into the world,

because there's literally no place to put it.

Like for me, if I have some reaction to something in the news,

I have to decide whether it's worth talking about it in my next podcast

that I might be recording, you know, four days from now.

And rather often people have been just bloviating about this thing for four solid days before I ever get to the microphone.

And then I get to think, what is, is it still worth talking about?

And most, almost nothing survives that test anymore, right?

So I could, the conversation has moved on.

So there's actually no place for me to just type this thing that either takes me 10 seconds and then rolls out there to get, to detonate in the minds of, you know,

my friends and enemies to opposite effect.

And then I see the, the result of all that, you know, on a, again,

on a, this sort of reinforcement loop of every 15 minutes.

Not having that is such a relief that I just don't even know why I would,

so like when threads was announced, I wasn't, I think I'm on threads too,

but it's not me.

It's just, you know, just get another marketing channel.

But yeah, I haven't, I feel such relief not exercising that muscle anymore,

where it's like, you know, I don't know how often I was checking Twitter,

but it was, I was, you know, I was not checking it just to see what was happening to me or the response to my last thing I tweeted.

I was checking it a lot because it was my newsfeed.

It's like I'm following, you know, 200 smart people.

They're telling me what they're paying attention to.

And so I'm fascinated.

So yeah, well, yeah, I want to see that next article or that next video,

just that engagement and the endless opportunity to comment and to put my foot in my mouth or put my foot in someone else's mouth or have someone put their foot.

It's just not having that has been such a relief that I would be,

I mean, it's not impossible, but I would be very cautious in reactivating that

because it was, it was so much noise.

And again, it created, there's so much, it became a, I mean, it became an opportunity cost, but it became a just this endless opportunity for misunderstanding,

but misunderstanding of me and, you know, everything I've been putting out into the world and then my sense that I had to react to it.

And then you just can't plow that back into the, you know,

that becomes the basis for further misunderstanding.

And it just constantly was giving me the sense that there's something,

there's something I need to react to on my podcast, in an article, on Twitter,

that it's just, this is a valid signal.

Like this is, this is, this is like, this is a five alarm fire.

This is like, you got to stop everything.

Like you're by the pool on the one vacation you're taking with your family that summer.

And this thing just happened on your phone that like, it can't wait, right?

Like you actually have to pay attention because it's like the conversation is happening right now.

And so it was a kind of addiction to information and, you know,

some level reputation management or, or, or, um,

I mean, just to just be free of it is, I mean, it's such a relief.

Apart from like, you know, health issues with certain family members,

virtually the only bad things that have happened to me have been a result of my engagement with Twitter

over the last 10 years.

So it's just, it's just, you know, I, you know, I guess I'm,

if I'm a masochist, I would be back on Twitter,

but like that would be the only reason to do it.

Narrow AI, I asked you the question a second ago, which we, um,

I really wanted to get a solution to it because I'm mildly terrified.

I completely believe your, believe your, um, the logic underneath your opinion that Narrow AI will cause this,

um, destabilization and unusability of the internet.

So just focusing on Narrow AI, what, what would you consider to be a solution to prevent us getting to that world

where misinformation is rife to the point that it can destabilize society, politics and culture?

Well, I think it's something I've been asking people about on my podcast because it's not actually my wheelhouse

and I would just need to hear from experts about what's possible technically here, but, um,

I'm imagining that paradoxically or ironically,

this could usher in a new kind of gatekeeping that we're going to rely on

because like the provenance of information is going to be so important.

I mean, the assurance that a video has not been manipulated

or there's not a, just a pure confection of, of deep fakery.

Right. So you get, so it could be that we're, we're meandering into a new period where you're not going to trust a photo unless it's come, it's coming from, you know, Getty images or, you know, the New York Times has some story about how they have verified every photo in there that they put in their newspaper. They have a process.

And, you know, so if you see a video of Vladimir Putin seeming to say that he's declaring war on the U.S.

Right. I think most people are going to assume that's fake until proven otherwise.

It's like it's just, it's just going to be too much fake stuff.

And it's going to be, it's going to all going to look so good that the New York Times and every other, you know,

organ of media that we have relied upon as imperfect as they've been of late,

they're going to have to figure out what the tools are whereby they can say,

okay, this is actually a video of Putin, right?

And if the new, I mean, I'm not going to be able to figure out on my own, right?

If the New York Times doesn't have a process or CNN doesn't have a process that they go through before they say,

okay, Putin really said this. And so this is, we have to now react to this because this is real.

Whatever that process is and whether it's, whether there's some kind of digital watermark that, you know,

that's connected to the blockchain, there's some tech implementation of it that can be fully democratized

where you, by just being in the latest version of the Chrome browser,

can know that you can differentiate, you know, real and fake videos.

I don't know what the implementation will be, but I just know we're going to get to some spot where it's going to be.

All right, we have to declare epistemological bankruptcy.

We don't know what's real. We have to assume anything, especially lured or agitating is fake until proven otherwise.

So prove otherwise. And that's, you know, that, that'll be a resetting of something.

I don't know what we do with that in a world where we really don't have that much time to react to certain things that are, you know,

a video of Putin saying he's launched his big missiles is something that, you know, 30 minutes from now.

we would, we would understand whether it's real or not.

We forget about, again, forget about everything we just said about AI.

Look at all of our legacy risks, look at the risk of nuclear war, the risk of stumbling into a nuclear war by accident has been hanging over our head for 70 years.

I mean, we've got this old tech, we've got these wonky radar systems that throw up errors.

We have moments in history where, you know, one Soviet sub commander decided based on his just gut feeling, his common sense that the data was almost certainly an error.

And he decided not to pass the obvious evidence of an American ICBM launch up the chain of command, knowing that the chain of command would say,

okay, you have to fire, right?

And he reasoned that if the U.S. was going to attack the Soviet Union, they would launch more than, I think in this case it looked like there were four missiles.

That was the radar signature.

If the U.S. is going to launch a first strike against the Soviet Union, this is like the mid-80s, they're going to launch more than four missiles.

This has to be bad data, right?

So if we automate all this, will we automate it to systems that have that kind of common sense, right?

But we've been perched on the edge of the abyss based on the possibility, forget about malevolent actors who might decide to have a nuclear war on purpose.

We have the possibility of accidental nuclear war.

You add this cacophony of misinformation and deep fake to all of that.

And it just gets scarier and scarier.

And this is not even AI.

This is just narrow AI amplified misinformation.

How do you feel about it?

Well, this is the thing that worries me.

I worry about the next election.

If we can run the 2024 election in a way that most of America acknowledges was valid, that will be an amazing victory, whatever the outcome.

I would not be looking forward to a Trump presidency, but I think even more fundamental than that is can we hold a presidential election 18 months from now

that is that we recognize as valid, right?

Like that.

I don't know.

I don't know what kind of resources are being spent on on that particular performance, but that is hugely important.

And I don't think our near term experiments with AI is going to make that easier.

Why is this important?

Well, it's just, I mean, if you think the maintenance of a valid democracy in the world's lone superpower is of minor importance.

I'd like to drink the tea you're drinking.

You're optimistic?

I mean, I can't say I'm optimistic.

It's a paradoxical state, I mean, because I definitely have.

I tend to focus on what's wrong or might be wrong.

I tend to, I think, have a pessimistic bias, right?

I tend to notice what's wrong as opposed to what's right.

That's my bias, but I'm actually very happy, right?

Like I have a very, a very good life.

I'm just like everything is just I'm incredibly lucky.

I'm surrounded by great people.

It's like it's all great.

And yet I see all of these risks on the horizon.

So I'm not, I just have a very high degree of well-being at this moment in my life.

And yet I like what's on the television is scary.

And so it's a very interesting juxtaposition.

I'll be very relieved if we have a, I feel like we're in a very weird spot.

I mean, like I haven't seen a full post-mortem on the COVID pandemic

that has fully encapsulated what I think happened to us there.

But my vague sense is that we didn't learn a whole hell of a lot.

I mean, basically what we learned is we're really bad at responding to this kind of thing.

This was a challenge that just fragmented us as a society.

It could have brought us together.

It didn't.

And it amplified all of the divisions in our society politically and economically and tribally and all kinds of ways.

The role of misinformation and disinformation and all of that was all too clear and I think just getting worse.

So I think as a dress rehearsal for some future pandemic that is inevitably going to come and could well be worse, I think we failed this dress rehearsal.

And I have to hope that at some point our institutions will reconstitute themselves

so as to be obviously trustworthy and engender the kind of trust we actually need to have on our institutions.

We need a CDC, not only that we trust, but that is trustworthy, that we're right to trust.

And so it is with an FDA and every other institution that is relevant here.

And we don't quite have that and half of our society thinks we don't have that at all.

And so we have to rebuild trust in institutions somehow and I just think we have a lot of work to do to even figure out how to make an increment of progress on that score.

Again, the siloing of large constituents into alternate information universes is just not functional.

And that's so much of what social media has done to us and alternative media.

You and I are podcasters, but I call it podcast to stand.

We have this landscape of now whatever, a million plus podcasts and there's email newsletters and everyone has now just decided to curate their information diet in a way that's just be poke to them.

And you can stay there forever and you're getting one slice and it could be a completely fictional slice of reality.

And we're losing the ability to converge on a common picture of what's going on.

So that sounds optimistic. I didn't hear the optimism in there.

You tell me.

No, but I kind of can't refute anything you said on a like a logical basis.

It all sounds like that is the direction of travel that we're going in, unfortunately.

I have faith that there'll be surprising positives.

There always tends to be surprising positives that we also didn't factor in.

It's easy to see if there's any significant low hanging fruit technologically or scientifically that could be AI enabled for us.

I mean, just take like a cure for cancer, a cure for Alzheimer's.

I mean, just having one thing like that.

That would be such an enormous good.

That's why we can't get off this ride.

And that's why there is no break to pull.

Because the value of intelligence is so enormous.

It's not everything.

There are other things we care about and a right to care about beyond intelligence.

Love is not the same thing as intelligence.

But intelligence is the thing that can safeguard everything you love.

Even if you think the whole point in life is to just get on a beach with your friends and your family and just hang out and enjoy the sunset.

You don't have to augment.

You don't need superhuman intelligence to do any of that.

You're fit to do it exactly as you are.

You could have done that in the 70s and it would just be just as good a beach.

It would be just as good friends.

But every gain we make in intelligence is the thing that safeguards that opportunity for you and everyone else.

I feel like we've not defined the term artificial general intelligence.

From my understanding of it, it's when the intelligence can think and make decisions almost like a human.

This is kind of just a semantic problem, but intelligence can mean many things.

But loosely speaking, it's the ability to solve problems and meet goals, make decisions in response to a changing environment, in response to data.

And the general aspect of that is an ability to do that in many different situations, all the sort of situations we encounter as people.

And to have one's capacity in one area, as I get better at deciding whether or not this is a cup, I don't magically get worse at deciding whether you just said a word.

I can do multiple things in multiple channels.

That's not something we had in our artificial systems for the longest time because everything was bespoke to the task.

We'd build a chess engine and it couldn't even play tic-tac-toe.

All it could do is play chess and we just would get better and better in these piecemeal, narrow ways.

And then things began to change a few years ago where you'd get, like, deep mind would have its algorithms that were the same algorithm with slightly different tuning could play go.

Or it could solve a protein folding problem as opposed to just playing chess.

And it became the best in the world at chess and it became the best in the world at go.

And amazingly, what AlphaZero did, before AlphaZero, all the chess algorithms were, they just had all of our chess knowledge plowed into them.

They had studied every human game of chess and it was a bespoke chess engine.

AlphaZero just played itself, I think for like four hours, right?

It just had the rules of chess and then it played itself and it became better not merely than every person who's ever played the game.

It became better than all the chess engines that had all of our chess knowledge plowed into them.

So it's a fundamentally new moment in how you build an intelligent system and it promises this possibility.

Again, this inevitability, the moment you admit that we will eventually get there.

The moment you admit that it can be done in silico and the moment that you admit that we will just keep going unless a catastrophe happens.

And those two things are so easy to admit that at this point I don't see any place to stand where you're not forced to admit them.

I don't see any neuroscientific or cognitive scientific argument for substrate dependence for intelligence, given what we've already built.

And again, we're going to keep going until something stops us.

We'll hit some immovable object that prevents us from releasing the next iPhone, but otherwise we're going to keep going.

And then, yeah, so then whatever general will mean in that first case, there'll be a case where we've built a system that is so good at everything we care about that is functionally general.

Now, maybe it's missing something.

Maybe it's missing something that we don't even have a name for.

We're missing all kinds of, there are possible intelligences that we haven't even thought about because we just haven't thought about them.

There are ways to section the universe undoubtedly that we can't even conceive of because we have the minds we have.

Elon was asked a question on this by a journalist.

The journalist said to him, in a world where you believe that to be true, that artificial general intelligence is around the corner,

when your kids come to you and say, Daddy, what should I do with my life to find purpose and meaning?

What advice do you now give them if you hold that intuition to be true?

That it's around the corner.

What do you say to your children when they say, what should I do with my life to create purpose and meaning?

Did you say that Elon answered this question?

What did he say?

It's one of the most chilling moments in an interview.

I've seen in recent times, because he stutters, he goes silent for about 15 seconds, which is very un-Elon.

He stutters, he stutters, he stutters a bit more.

And then he says, he thinks he's living in suspended disbelief because if he really thought about it too much, what's the point?

He says, what's the point of me building all these cars?

He was in his Tesla factory.

What's the point of me building all these cars and what's the point?

I do think that sometimes.

So I think I have to live in, as his words were, suspended disbelief.

Right.

Well, I would encourage him to ask what's the point of spending so much time on Twitter, because he could clearly benefit from rethinking that.

But that aside, I mean, my answer to that is, and I think other people have echoed this of late, I mean, sort of surprising to me.

My answer is that this begins to privilege a return to the humanities as a kind of a core, like the center of mass intellectually for us.

Because when you look at what we're really good at, and it's among the last things that can be plausibly automated.

And if we automate it, we may cease to care about it.

So it's like learning to write good code is something that is being automated now.

I'm not a programmer, but I have it on good authority that already these large language models are improving code and something like half the time they're writing better code than people.

But that's all going to become like chess, right?

It's just it's going to be better than people ultimately.

So being a software engineer is something that, you know, and being a radiologist and being like those things.

It's easy to see how AI just cancels those professions or at least makes one person, you know, so effective at using AI tools that, you know, one person can do the work of 100 people.

Sometimes people who don't have to be doing that job.

But creating art and, you know, writing novels and being a philosopher and talking about what it

means to live a good life and how to do it.

That's something that if we have to look at where we're going to care that we're actually in relationship to and in dialogue with another person who we know to be conscious.

Where we don't care about that, we're not going to care.

We're going to want just the best version of it.

The cure for cancer comes from an AI, an incentive in AI.

I do not give a shit.

I just want the cure for cancer, right?

Like there's no added value that where I find out, okay, the person who gave me this cure really felt good about it.

And he's, you know, he had tears in his eyes when he figured out the cure.

Every engineering problem is like that.

We want safer planes.

We want, you know, we just want things to work.

We're not sentimental about the artistry that went into all of that.

And when the difference, when the gulf between the best and the mediocre gets big and consequential, we're just going to want the best.

We're just going to want the best all the way down the line.

But what is the best novel, right?

What is the best podcast conversation?

And can you subtract out the conscious person from that and still think it's the best?

And so like someone once sent me what reported to be, I didn't even listen to it, so I'm not even sure what it was.

But it looked like it was an AI-generated conversation between Alan Watts and Terrence McKenna, right?

Both guys who I love.

I remember I didn't know either of them, but some fans of both have listened to hundreds of hours of both talk.

As far as I know, they never met each other.

It would have been a fascinating conversation.

I realized when I looked at this YouTube video, I realized I simply don't care how good this is because I only care if it was actually Alan Watts and Terrence McKenna talking.

A simulacrum of Alan Watts and Terrence McKenna in this context, I don't care about.

So in another use case, I stumbled upon, I was playing with ChatGPT and I asked the causes of World War II,

give me 500 words on the cause of World War II, and it gives you this perfect little bullet-pointed essay on the cause of World War II.

That's exactly what I want from it.

That's fine.

I don't care that there was no person behind that typing.

But when I think, well, do I want to read Churchill's history of World War II?

It's on my shelf to read.

It's one of these aspirational sets of books.

I haven't read it yet.

I actually want to read it because Churchill wrote it.

And if you could give me an AI version of Churchill saying this is in the style of Churchill, even Churchill scholars say this sounds like Churchill, I actually don't care about it.

That's not the use.

I'll take the generic use of give me the cause of World War II.

The fake Churchill is profoundly uninteresting to me.

The real Churchill, even though he's dead, is interesting to me.

The rebuttal I give here, and this is what my mind is doing, is saying the distinction you're presenting,

the difference I see is that in the case of the conversation between two people you respect that has been generated by AI,

someone has signaled to you that it is fake.

If you remove that because say Churchill thought, why would I write a book when I could just click a button

and this thing will write it in my voice, in my tone of voice, with the entire back catalogue of things I've written before.

And it will produce my account and it will save me time.

So I'll just click a button and my publisher maybe will do it for me.

And then I'll sell that to Sam on the basis that it is my thoughts, which I can imagine a very near future.

If we just do it by percentage, how many books are going to be increasingly written by artificial intelligence?

To the point that when you look at a shelf, I imagine at some point in the future,

if the intelligence does increase by any measure, that most of it would be words strung together by artificial intelligence

and it will be selling potentially better than the words written by humans.

So again, when we go back to the conversation with your children, there might not be a career there either

because artificial intelligence is faster, can produce more, can test and iterate on whether it sells better,

clicks gets more clicks, it can write the headline, create the picture, write the content,

and then I can just take the check because I put my name to it.

So even in that regard, what remains?

Well, so in the limit, what I think we're imagining is a world where,

and so none of the terrifyingly bad things have happened, so it's just all working.

We're just producing a ton of great stuff that is better than the human stuff and people are losing their jobs.

We've got a labor disruption, but we're not talking about any other kind of political catastrophe or cyber apocalypse, much less AGI destroying everything.

Then I think we just need a different economic assumption and ethical intuition around the value of work.

Our default norm now in a capitalist society is you have to figure out something to do with most of

your time

that other people are willing to pay you for.

You have to figure out how to add value to other people's lives such that you reliably get paid.

Otherwise, you might die.

We've got a social safety net, but it's pretty meager.

There are cracks you can fall through.

You could wind up homeless, and we're not going to figure out what to do about that.

We're all too well.

So your claim upon your existence among us is you finding something to do with your time that other people will pay you for.

And now we've got artificial intelligence removing some of those opportunities, creating others.

But in the limit, and I do think it is different.

I think analogies to other moments in technological history are fundamentally flawed.

I think this is a technology which in the limit will replace jobs and not create better new jobs in their wake

This just cancels the need for human labor ultimately.

And strangely, it replaces some of the highest status most cognitively intensive jobs first.

It replaces Elon Musk before it replaces your electrician or your plumber or your masseuse way before.

So we have to internalize the reality of that.

Again, this is in success.

This is all good things happening.

And we have to have a new ethic.

We have to have a new economics based on that ethic, which is UBI is one solution to this.

You shouldn't have to work to survive.

Universal basic income.

Yeah, there's so much abundance now being created.

We have to figure out how to spread this wealth around.

We've got a cure for cancer over here.

We've got perfect photovoltaic driven economies over here.

We've solved the climate change issue.

We're just pulling wealth out of the ether, essentially.

We've got nanotechnology that is just birthing whole new industries yet, but it's all being driven by AI.

There's no room in this.

Whenever you put a person in the decision chain, you're just adding noise.

This is the best thing.

This should be the best thing that's ever happened to us.

This is just like God handing us the perfect labor saving device.

The machine that can build every other machine that can do anything you could possibly want.

We should figure out how to spread the wealth around in that case.

This is just powered by sunlight, no more wars over resource extraction.

It can build anything.

We can all be on the beach just hanging out with our friends and family.

Did you believe we should do universal basic income where everybody's given a monthly check? We have to break this connection.

Again, this is what will have to happen in the presence of this kind of labor force dislocation enabled by all of this going perfectly well.

This is pure success.

AI is just producing good things.

The only bad thing is putting all these people out of work.

It's coming for your job eventually.

My issue with it, and my rebuttal when I talked to my friends about this idea of universal basic income

when we hand out enough cash or resources to people so that they're stable,

which I'm not necessarily against, but just want to play with it a little bit,

is humans seem to have an innate desire for purpose and meaning,

and we seem to be designed and built psychologically for labor and for discomfort.

But it doesn't have to be labor that's tied to money.

We will get our status in other ways and we'll get our meaning in other ways.

Again, these are all just stories we tell ourselves.

You're talking to a person who knows it's possible to be happy actually doing nothing,

like just sitting in a room for a month and just staring at the wall.

Because you've done it.

Like that's possible.

Yet that's most people's worst nightmare.

It's the solitary confinement in a prison that's considered a torture.

And I know people who spent 20 years in a cave.

So it's like there are capacities here that are worth talking about.

But just more commonly, I think we want to be entertained.

We want to have fun.

We want to be with the people we love.

We want to be useful in relationship.

And insofar as that gets uncoupled from the necessity of working to survive.

It doesn't all just go away.

We just need new norms and new ethics and new conversations around what we do on vacation.

What you're imagining is that if you put everyone on vacation, on the best vacation,

if you take the vacation as good as possible,

a majority of people will eventually be miserable because they're not back at work.

And yet most of these people are working so that they have enough money

so that they could finally take that vacation.

We will figure out a new way to be happy on the beach.

If you get bored with Frisbee, we will figure something else out that is fun.

I'll be able to read the Churchill history of World War II on the beach

and not be rushed by any other imperative because I'm happily retired

because my AI is creating the thing that is solving all my economic problems.

We should be so lucky as to have that be our problem.

How to be happy in conditions of no economic imperative,

no basis for political strife on the basis of scarce resources,

and the question of survival is off the table with respect to what one does with one's time and attention.

You can be as lazy as you want and you'll still survive.

The awful situation we're in now is that differences in luck mean everything.

Someone is born without any of the advantages that we have.

We don't have an economic system that reliably gives them every advantage and opportunity they could have.

We've convinced ourselves we either don't have the resources or we've convinced ourselves we don't have the resources.

We don't have the incentive such that we access the resources

so as to actually come to the help of people we could help.

The idea that people starve to death is just unimaginable and yet it still happens.

That's not a scary thing to do.

It's a political problem wherever it happens.

Yet all of this is tied to a system where everyone has convinced themselves that it's normal to really have one's survival be in question if one doesn't work.

By choice or by accident, if you haven't, I think,

at least in the U.S., this is almost certainly not true in the U.K.,

but in the U.S., the most common reason for a personal bankruptcy is overwhelming medical expense

that just comes upon you for whatever reason.

Your wife gets cancer, you guys go bankrupt solving the cancer problem

or failing to solve the cancer problem and now everything else is going to be fine.

We have a society which thinks if you wind up homeless,

just don't sleep in front of my store because you're going to hurt my business.

Successful AI that cancels lots of jobs would only be cancelling those jobs

by virtue of producing so many good things, so much value for everybody

that we would have to figure out how to spread that wealth around.

Otherwise, we would have an amazingly dystopian bottleneck for a few short years

and then we would just have a revolution.

Then the guys in their gated communities making trillions of dollars

based on them having gotten close enough to the GPUs that some of it rubbed off on them,

they'd be dragged out of their houses and off their Gulf Streams

and we would have a fundamental reset, a hard reset of the political system.

If I had to put you in a yes or no situation and ask your intuition the question now

that if your objective was to, which I'm sure it is,

is to encourage the betterment of humanity

and to increase our odds of happiness and well-being 100 years from now

and there was a button placed in front of you

and it would either end the development of artificial intelligence

as we've seen it over the last decade,

so it would never proceed with developing intelligent machines.

Or not, so you could press a button and stop it right now.

And stop it permanently such that we never then do that thing?

We just never figure out how to build intelligent machines?

Pause it indefinitely.

Well, I would definitely pause it to a point where we would get our heads around the alignment problem.

Permanently, if the button was a permanent pause that you couldn't undo.

Well, the question is how deep does that go?

So like we have everything we have now, but we just never get better than now?

Yeah, we never make progress from here.

And your objective is to make humanity happy and prosperous?

It's hard because when you begin imagining all of the good stuff

that we could get with aligned superhuman AI,

then it's just cornucopia upon cornucopia.

It's just everything is potentially within reach.

Yeah, I mean, I take the existential risk scenario seriously enough that I would pause it.

I think we will eventually get to it.

If curing cancer is a biomedical engineering problem that admits of a solution,

and I think there's every reason to believe it ultimately would be,

we will eventually get there based on our own muddling along with our current level of tech, currently information tech.

I'm reasonably confident of that.

Our intelligence shows every sign of being general is just it's not as fast as we would want it to be.

The thing that AI is going to give us is it's going to give us speed that is...

There's speed and then there's the access, there's memory.

We can't integrate. We don't have the ability.

No person or team of people can integrate all of the data we already have.

The real promise here is that these systems will be able to find patterns

that we wouldn't even know how to look for and then do something on the basis of those patterns.

I think an intelligent search within the data space by apes like ourselves

will eventually do most of the great things we want done.

The problems we need to solve so as to safeguard the career of our species

and to make civilization durable and sane and to remove this sort of Damocles

that is over our heads at every moment that at any moment we could just decide to have a nuclear war

that ruins everything or create an engineered pandemic that ruins everything.

We don't need superhuman intelligence to solve all those problems.

We need an appropriate emotional response to the untenability of the status quo

and we need a political dialogue that eventually transcends our tribalism.

For those of you that don't know, this podcast is sponsored by Woop, a company that I'm a shareholder in

and I'm obsessed with my Woop. It's glued to my wrist 24-7

and for those of you that don't know, it's essentially a personalized wearable health and fitness coach

that helps me to have the best possible health.

My Woop has literally changed my life.

Woop is doing something this month which I'd highly suggest checking out.

It's a global community challenge called the Core 4 Challenge.

Essentially, they guide you through a set of four activities throughout the month of August

that are scientifically proven to improve your overall health.

I'm giving it a go and I can't wait to see the impact it has on me

and I highly recommend you to join me with that.

So if you're not on Woop here, there is no better time to start.

If you're a friend of mine, there's a high probability that I've already given you a Woop

because I'm that obsessed with it.

This is the thing that I check when I wake up in the morning.

It's the first thing that I look at.

I want the information on my sleep to then plan my day around.

So if you haven't joined Woop yet, head to join.woop.com.co

to get your free Woop device and your first month free.

Try it for free and if you don't like it after 29 days,

they're going to give you your money back but I have a suspicion that you're going to keep it.

Check it out now and let me know how you get on.

Send me a DM.

Quick one.

If you've been listening to this podcast for some time,

I've been sharing messages you've heard over and over and over again,

especially when we first had that conversation with Tim Spector,

is about the importance of greens in our diet.

And a while ago, I started pressing my friends at Huell

to come out with a product that did exactly that,

allowed you to have all those greens, the vitamins and minerals you need in a drink.

And after several, several months of iterations and processes,

they released this product called Huell Daily Greens,

which is now one of my favorite products from Huell because it tastes great

and it fills that very important nutritional gap that I had in my diet.

The problem is, it launched in the US and it sold out straight away

and became a smash hit for Huell for the very reasons I've described.

It's now back in stock in the United States, but it's not here in the UK yet.

So if you're a UK listener, which I know a lot of you are, it's not yet available.

So let's all attack Huell.

Let's DM them everywhere we can and tell them to bring Huell Daily Greens to the UK.

This is the product. When it is available in the UK, I'm going to let you know first,

but until then, let's spam their DMs.

You and, I'd say a few others, maybe two or three others helped change my mind about one of the most profound things I think anyone could believe,

which was when I was 18, I believed in Christianity.

And then there was a couple of moments that shook my belief.

Nothing on a personal level, just a couple of ideas that managed to sort of infect my operating system

that led my curiosity towards your work.

And I changed my mind profoundly.

It's such a profound change that I had.

How do we change our minds?

And I really want to focus that question on the individual's mind.

I want to change my mind.

I want better beliefs, better ideas in my head that are going to allow me to get out of my own way.

Because I am not a cheat. I'm miserable.

I'm not living the life that I...

I would say I know I can live, but some people don't even know they can live a better life.

I'm not happy. That's the signal.

And I want to rectify this in some way.

Yeah, well, there are a few bright lines for me.

I mean, we sort of take our ethical lives and our relationships to other people, right?

So there's the problem of individual well-being that's kind of still real,

even if you're in a moral solitude, if you're on a desert island by yourself,

you really don't have ethical questions that are emerging because you're not in a relationship with anybody else,

but you still have the problem of how to be happy.

But so much of our unhappiness is in collaboration with others, right?

We're unhappy in our relationships. We're unhappy professionally.

And it's worth looking at how we're behaving with other people.

For me, the highest leverage change I ever made, and it's, again, it's very easy to spell out, and it's very clear.

And ultimately, it's pretty easy is just to decide that you're not going to lie about anything, really.

I mean, there might be some situations in extremis where you'll feel forced to lie,

but those, in my view, are analogous to acts of violence that you may be forced to use in self-defense, right?

So like, lying is sort of the first stage on the continuum of violence for me, right?

So like, I'm not going to lie to someone unless I recognize that this is not a rational actor who I can possibly collaborate with.

This is someone I have to be, I have to avoid or defeat or otherwise contain their propensity to do me harm.

So, yes, if the Nazis come to the door and ask if you've got Anne Frank in the attic, yes, you can lie, or you can shoot them, or you can, these are not normal circumstances.

But that aside, every other moment in life where people are tempted to lie is one that I think you can categorically rule out as being unethical

and beyond unethical, it's just not, it's creating a life you don't, when you examine it, you don't want to live, right?

I mean, the moment you know that you're not going to lie to people and they know that about you, it's like all of the dials get, the social dials get sort of recalibrated on both sides,

and then you find yourself in the presence of people who don't ask you for your opinion unless they really want it, right?

And then when you're honest, I mean, then it's a night and day difference when you're giving people feedback, critical feedback,

and they know you're honest, right? They know their bullshit detector is not going off because they just know you're,

even when it's not convenient, you're being honest, or even when it's not comfortable, you're being honest.

One that's incredibly valuable because basically you're giving them the information that you would want if you were in their shoes, right?

Because we have this sort of delusion that takes over us whenever we're tempted to tell a white lie.

We imagine, okay, this person doesn't want, it'd be much better for me to just tell them the kind fiction than tell them the uncomfortable truth, right?

But we don't do the, so we don't even calculate for the golden rule there most of the time, and we, if you just took a moment, you'd realize,

well, wait a minute, does someone who is actually doing a bad job want me to tell them that they're doing a good job and then just send them out into the world

to bounce around other people who are going to be recognizing, as I just did, that the thing they're doing isn't so great, right?

You're just not doing them a favor, right?

This is part of the nature of belief change, isn't it?

That when someone, we believe that someone is on our side, or we believe from like a political standpoint that they represent, 99% of the views that we represent,

we're much more likely to change our beliefs.

I spoke to Tali Sharra about this, the neuroscientist, and I wrote about this in a chapter in my upcoming book about how you change people's minds.

They showed in the elections that if like a flat earther says something to a flat earther about the nature of the earth, or believe it, but if NASA says something to a flat earther,

they will just dismiss it on site because the source of that information is not one that they believe, or trust, or like, or believe is well-intentioned.

I mean, this is a bug, not a feature.

I mean, it's understandable, but this is something we have to grow beyond because the truth is the truth, right?

So you can't, I mean, it goes in both directions.

The person on your team who you love and respect is capable of, in their very next sentence, of speaking of falsehood, right?

And you need to be able to detect that.

And conversely, the person you least respect is capable of saying something that's quite incisive and worth taking on board.

And so we have to have this sort of metacognitive layer where we're noticing how we're getting played by our social alliances and recognize that the truth,

and rather often important truths, are evaluated by different principles.

I mean, it's not a matter of the messenger, you know, you shouldn't shoot the messenger and you shouldn't worship him.

You mentioned lying as being a, well, removing lying and being more honest is being a significant step change in your own happiness.

Is that accurate?

In my happiness.

In your own happiness?

Yeah, yeah, yeah.

Yeah, immensely so.

Practically and specifically how?

And when you look at how people ruin their reputations and their relationships and their businesses, their careers, the gateway to all of the misbehavior that accomplishes that is lying.

It's, I mean, you look at somebody that lands arms strong, right?

I mean, just, or Tiger Woods, right?

These guys are the absolute apogy of sport.

Everyone loves them.

Everyone's just amazed at what they've accomplished.

And yet, you know, the dysfunction in their lives just gets vomited up for all to see at a certain point. And it was just enabled at every stage along the way by lying.

If either of them had early in their career before they became famous, before they became rich, before they became tempted to do anything that was going to derail their lives later on, if they had decided they weren't going to lie, right?

They would have found all everything else they did to screw up their success impossible.

So when I decided, and this is in the book, this is a course I took at Stanford.

It was a seminar with this brilliant professor, Ron Howard, who many people, some people in Silicon Valley have taken his course as well.

I mean, this course was just like a machine.

Undergraduates and graduate students would come in on one side, and then 12 weeks later would come out convinced that basically lying was no longer on the menu, right?

It was that the whole seminar was an analysis of the question, is it ever right to lie?

And really we focused on white lies, truly tempting lies as opposed to the obvious lies that screw people's lives and relationships.

It's just so corrosive, and it's corrosive of relationships in ways that you, unless you're a student of this kind of thing, you don't necessarily notice.

I mean, one example I believe is in that book is that I remember my wife was with a friend, and the two of them were out, and the friend had something she had to do with another friend later that night, but she didn't really feel like doing it.

And she got a call from that friend in the presence of my wife, and she just lied to the friend to get out of the plan, right?

She said, oh, you know, I'm so sorry, but my daughter's got this thing, and it was just an utterly

facile use of dishonesty.

She could have just been honest, right, but it was just too awkward to be honest, so she just got out of it with a lie.

But now it's in the presence of my wife, and my wife is now, the immediate question is, how many times have I been on the other side of that conversation, right?

How many times has she lied to me in an equally compelling way about something so trivial, right? And so it just eroded trust in that relationship in a way that the liar would never have known about, would never have detected it, because it's just she just went right back to having a good time with, you know, they were just out to lunch,

and they continued, you know, having their lunch, and they're still having a good time, and it's all smiles, but my wife has just logged something about kind of the ethical limitations of this person, and the person doesn't know it, right?

And so once you sort of pull on this thread, you basically, your entire life becomes, at least for the transition period, until this just becomes a habit you no longer have to consider,

suddenly it's your, the world becomes kind of mirror thrown up to your mind, and you meet yourself in all these situations where you were avoiding yourself before.

So like someone will say, you know, do you want to have plans, or do you want to collaborate with me on this project?

And if previously you always had recourse to some kind of white lie that just got you out of, you know, the awkward truth, which is the answer is no, and there are actually reasons why not, right? You never have to confront the awkwardness of that, you're this kind of person who has these kinds of commitments, and this kind of, you know, it's like, you know, I mean the most awkward one would be, you know, someone declares a romantic interest in you,

and the answer is no, and then the, it's no for a totally superficial reason, right? Like this person is, they're not attractive enough for you, right? Or they're overweight, or what it is, it's just like you have your reason why not, and this is something you feel you cannot say, right?

Now I'm not saying that you should always go out of your way like someone with Tourette's who just helplessly blurts out the truth, like there's a scope for kindness and compassion and tact, but if someone is going to really drill down on the reasons why not,

if the person says, no, I want to know exactly why you don't want to go out with me, there's something to discover on either side of that true disclosure, right? Like either you are cast back on yourself and you have to realize, okay, I'm such a superficial person that it doesn't matter who anyone is,

if they're 10 pounds overweight, I'm not interested, right? That's, that's the mirror held up to your mind, it's like, okay, all right, so you're that kind of person, do you want to still be that kind of person? Do you really want to just decide that everyone, no matter what their virtues, right? And no matter what has been going, you know,

no matter what chaos is going on in their life and they actually, this person might actually lose those 10 pounds next month and you would have a very different situation, but are you really not available, are you really filtering by weight in this way?

And are you really comfortable with that? And are you comfortable saying that? Like if somebody forces you to actually be honest?

We have a closing tradition on this podcast where the last guest leaves a question for the next guest,

not knowing who they're going to leave it for. The question that's been left for you.

Peckable handwriting. Where do you want to be when you die? Describe the place, time, people, smell and feeling?

Well, this actually connects with an idea I've had. I mean, I think what we need, we haven't talked about psychedelics here, but there's been this renaissance in research and psychedelics and it's hard to know.

I'm worried that we could recapitulate some of the errors of the 60s and roll this all out in a way that's less than wise, but the wise version would be, I think we need to recapitulate something like the Mysteries of Elusis where we have rites of passage that are enabled by, in many people's case, psychedelics and the practice of meditation.

I just think these are just fundamental tools of insight that are, for most people, it's hard to see how they would get them any other way. There's a longer conversation about which molecule and how and all that, but another component of this is a hospice situation where the experience of dying is as wisely

embraced and facilitated as is possible. And I think psychedelics could certainly play a role for many people there. So I imagine something like we need places that are truly beautiful where people have gone to die and their families can visit them there.

And it is just a final rite of passage that is embraced with all the wisdom we can muster there. So in my case, I would want to be in, currently I'd be happy to be home, but wherever home is at that point, I would want a view of the sky.

It could be an ocean beneath the sky, that would be ideal. There's basically nothing that makes me happier than just looking at a blue sky with just watching cumulus clouds move across a blue sky. I can extract so much mental pleasure just looking at that. So yeah, if I'm going to spend my last hours of life looking at anything, if my eyes are going to be open, looking at the sky and having... The stars with the sky, the daytime sky. The sky, the daytime. Light pollution is enough of a thing in my world that I feel like I go for years without seeing a good night sky. So I've kind of given up hope there, but I do love that.

But yeah, just a view of the sky and with the people I love at that point who are still alive at that point. Yeah, I'm not worried about death in that sense. I really think it's... The death part is not a problem.

I can't say I'm looking for... I can imagine there could be sort of medical chaos and uncertainty and all of the weirdness that happens around the dying process, right, depending on...

And there are all kinds of ways to die that I wouldn't choose, but having a nice place to do that with a view of the sky would be the only solution I think I would require.

The question asks the smell.

Give me the smell. Give me an ocean breeze. I have put an ocean there, so yeah, an ocean breeze would be perfect.

Sam, thank you so much. Thank you for not just this conversation. As I said to you before you sat down, you were pivotal in really helping me to unpack some problems on our younger, some conflicts I should describe them as with my view on religious belief and the nature of the world. More importantly, you didn't rob me of my religious beliefs and leave me with nothing. You left me with something else, which is something that was really important to me, which was the idea that there can still be great meaning and there can be what you describe as spirituality in the absence or

in the place of that religious belief.

Religious belief gives people a lot of things and it's funny because when I was religious and I went on the journey to becoming agnostic, let's say, I was in conflict with people.

As in, I would want to have a debate with everybody and I spent those two years watching everything that you and Richard Dawkins and Hitchens had all done and then I came out of the side and it was peaceful.

And it's you believe what you want, I'll believe what I want. As long as we're not causing any conflicts with each other and you're not doing any harm, it's okay.

And then I discovered what I would call my own spirituality, which is my meaning, the meaning that I see in the world around me and the self and things like psychedelics.

And it's a better place to be and it removed my fear of death, which I had as a religious person. So thank you. Thank you for that.

And all your subsequent work, incredible books, you've written so many of them that are absolutely incredible.

You've got an unbelievable podcast, which I was gorging on before you came here as well in an app, which I mean, if you could speak just a few sentences about the meaning of the app and what you do.

I know it's much more than meditation now, but I think people listening to this might be compelled to check it out and download it.

Yeah, well, so I had that book, which you're holding and waking up, which is the, which is where I talk about my experience in meditation and just how I fit it into a scientific secular worldview.

And just it just turns out that an app is a much better delivery system for that kind of information.

I mean, it's just hearing audio is you don't even need you don't need video.

I think audio is the perfect medium for it.

So when that technology came about or when I discovered it, I just felt incredibly lucky to be able to build it.

And so it's it's kind of outgrown me now.

There are many, many teachers on it and many other topics beyond meditation that are touched. But it's a it really subverts all of the problems that, you know, some of which we touched upon here with the with the smartphone.

I mean, like the smartphone has become this this tool of fragmentation for us.

It fragments our attention.

It continually interrupts our experience.

It's depending on how you use it.

But most of what we do with it, you know, you're checking Slack, you're checking your email, you're checking your social media, you're, you're just it's punctuating your life with all these, you know, at this point, seemingly necessary interruptions.

But this app or, you know, any really any app like it that is delivering this kind of content subverts all that because it's just this is this is it's just a platform where you're getting audio that is guiding you in a specific very specific use of attention and a sort of reordering of your priorities and getting you to recognize things about your experience that you wouldn't otherwise see.

And yeah, an app is it just a sheer good luck.

It turns out it's just the perfect delivery system for that information.

So yeah, I just felt very lucky to have stumbled upon it because again, you know, 10 years ago, there were no apps and you know, there's just it's just all I could do is write a book.

Sam, thank you.

Yeah, thank you.

Thank you so much.

Pleasure to meet you.

Yeah, pleasure to meet you.

Congratulations with everything.

It's really I was catching up on your podcast in anticipation of this.

And it's amazing the reach you've got now.

So yeah, it's wonderful.

No, it's still trying to catch up with it, but it's a credit to all of the team.

And I really want to say from the bottom of my heart, thank you because the work you do is really, really important.

It's been important in my life, as I've said, but it's just really important.

And I feel like we're living in a world where like nuance and all the things you've talked about and openness to debate and honest dialogue us, we're getting further and further away from there.

So if there's anyone left in this world that's still willing to engage on that level, I feel like they must be protected at all costs and I see you as one of those people.

So thank you.

Nice.

Nice.

Well, to be continued.

Yeah.