Are you uncomfortable talking about this?

Yeah, I mean it's pretty wild, right?

I'm going to start by sullying mine.

The billionaire founder of Google's AI technology.

He's played a key role in the development of AI from its first critical steps.

In 2020 I moved to work on Google's chatbot.

It was the ultimate technology.

We can use them to turbocharge our knowledge unlike anything else.

Why didn't they release it?

We were nervous. We were nervous.

Every organisation is going to race to get their hands on intelligence.

And that's going to be incredibly destructive.

This technology can be used to identify cancerous tumours

as it can to identify a target on the battlefield.

A tiny group of people who wish to cause harm

are going to have access to tools that can instantly destabilise our world.

That's the challenge how to stop something that can cause harm and potentially kill.

That's where we need containment.

Do you think that it is containable?

It has to be possible.

Why?

It must be possible.

Why must it be?

Because otherwise it contains us.

Yet you chose to build a company in this space.

Why did you do that?

Because I want to design an AI that's on your side.

I honestly think that if we succeed, everything is a lot cheaper.

It's going to power new forms of transportation, reduce the cost of healthcare.

But what if we fail?

The really painful answer to that guestion is that...

Do you ever get sad about it?

Yeah.

It's intense.

Better and better and better.

I can't tell you how much it helps when you hit that subscribe button.

The show gets bigger, which means we can expand the production,

bring in all the guests you want to see and continue to do this thing we love.

If you could do me that small favour and hit the follow button,

wherever you're listening to this, that would mean the world to me.

That is the only favour I will ever ask you.

Thank you so much for your time.

Everything that's going on with artificial intelligence now

and this new wave and all these times like AGI

and I saw another time in your book called ACI.

It's the first time I'd heard that term.

How do you feel about it emotionally?

If you had to encapsulate how you feel emotionally about what's going on in this moment, what words would you use?

I would say in the past it would have been petrified.

And I think that over time, as you really think through the consequences

and the pros and cons and the trajectory that we're on,

you adapt and you understand that actually there is something incredibly inevitable

about this trajectory and that we have to wrap our arms around it

and guide it and control it as a collective species, as humanity.

And I think the more you realise how much influence we collectively can have over this outcome,

the more empowering it is because on the face of it,

this is really going to be the tool that helps us tackle

all the challenges that we're facing as a species.

We need to fix water desalination.

We need to grow food 100x cheaper than we currently do.

We need renewable energy to be ubiquitous and everywhere in our lives.

We need to adapt to climate change.

Everywhere you look, in the next 50 years, we have to do more with less

and there are very, very few proposals,

let alone practical solutions for how we get there.

Training machines to help us as aides, scientific research partners, inventors, creators

is absolutely essential and so the upside is phenomenal, it's enormous.

But AI isn't just a thing, it's not an inevitable whole.

Its form isn't inevitable, right?

It's form, the exact way that it manifests and appears in our everyday lives

and the way that it's governed and who it's owned by and how it's trained.

That is a question that is up to us collectively as a species to figure out over the next decade

because if we don't embrace that challenge, then it happens to us.

And that's really what I have been wrestling with for 15 years of my career

and how to intervene in a way that this really does benefit everybody

and those benefits far, far outweigh the potential risks.

What stage were you petrified?

So I founded DeepMind in 2010

and over the course of the first few years, our progress was fairly modest.

But quite quickly in sort of 2013 as the deep learning revolution began to take off,

I could see glimmers of very early versions of AIs learning to do really clever things.

So for example, one of our big initial achievements was to teach an AI to play the Atari games.

So remember Space Invaders and Pong where you bat a ball from left to right.

And we trained this initial AI to purely look at the raw pixels screen by screen, flickering or moving in front of the AI

and then control the actions up, down, left, right, shoot or not.

And it got so good at learning to play this simple game simply through attaching a value between the reward

like it was getting score and taking an action

that it learned some really clever strategies to play the game really well

that us games players and humans hadn't really even noticed, at least people in the office hadn't noticed it, some professionals did.

And that was amazing to me because I was like, wow, this simple system

that learns through a set of stimuli plus a reward to take some actions

can actually discover many strategies, clever tricks to play the game well that us humans haven't occurred to us, right?

And that to me is both thrilling because it presents the opportunity to invent new knowledge and advance our civilization.

But of course, in the same measure is also petrifying.

Was there a particular moment when you were at DeepMind where you had that kind of eureka moment like a day

when something happened and it caused that epiphany, I guess?

Yeah, it was actually a moment even before 2013 where I remember standing in the office and watching a very early prototype of one of these image generation models

that was trained to generate new handwritten black and white digits.

So imagine zero to one, two, three, four, five, six, seven, eight, nine, all in different style of handwriting

on a tiny grid of like 300 pixels by 300 pixels in black and white.

And we were trying to train the AI to generate a new version of one of those digits, a number seven in a new handwriting.

Sounds so simplistic today given the incredible photorealistic images that are being generated, right?

And I just remember so clearly it took sort of 10 or 15 seconds and it just resolved that the number appeared.

It went from complete black to like slowly gray and then suddenly these white pixels appeared out of the black darkness

and it revealed a number seven.

And that sounds so simplistic in hindsight, but it was amazing.

I was like, wow, the model kind of understands the representation of a seven well enough to generate a new example of a number seven,

an image of a number seven, you know, and you roll forward 10 years and our predictions were correct.

In fact, it was quite predictable in hindsight, the trajectory that we were on.

More compute plus vast amounts of data has enabled us within a decade to go from predicting black and white digits,

generating new versions of those images to now generating unbelievable photorealistic, not just images,

but videos, novel videos with a simple natural language instruction or a prompt.

What has surprised you? You said you referred to that as predictable, but what has surprised you about what's happened over the last decade?

So I think what was predictable to me back then was the generation of images and of audio because the structure of an image is locally contained.

So pixels that are near one another create straight lines and edges and corners,

and then eventually they create eyebrows and noses and eyes and faces and entire scenes.

And I could just intuitively, in a very simplistic way, I could get my head around the fact that, OK, well, we're predicting these number sevens.

You can imagine how you then can expand that out to entire images, maybe even to videos, maybe, you know, to audio too.

What I said, you know, a couple of seconds ago is connected in phoneme space in the spectrogram. But what was much more surprising to me was that those same methods for generation applied in the space of language.

You know, language seems like such a different abstract space of ideas.

When I say like the cat sat on the most people would probably predict Matt, right?

But it could be table, car, chair, tree, it could be mountain, cloud.

I mean, there's a gazillion possible next word predictions.

And so the space is so much larger.

The ideas are so much more abstract.

I just couldn't wrap my intuition around the idea that we would be able to create the incredible large language models that you see today.

You chat GPTs.

Chat GPT.

The Google's Bard.

Inflection, my new company has an AI called Pi, a Pi.ai, which stands for personal intelligence.

And it's as good as chat GPT, but much more emotional and empathetic and kind.

It's just super surprising to me that just growing the size of these large language models as we have done by 10x every single year for the last 10 years, we've been able to produce this.

And that's just an amazingly large number.

If you just kind of pause for a moment to grapple with the numbers here.

In 2013, when we trained the Atari AI that I mentioned to you at DeepMind, that used two PETA flops of computation.

So PETA PETA stands for a million billion calculations.

A flop is a calculation.

So 2 million billion, right?

Which is already an insane number of calculations.

Lost me at two.

It's totally crazy.

Yeah, just two of these units that are already really large.

And every year since then, we've 10x'd the number of calculations that can be done such that today, the biggest language model that we train at Inflection uses 10 billion PETA flops.

So 10 billion million billion calculations.

I mean, it's just unfathomably large number.

And what we've really observed is that scaling these models by 10x every single year produces this magical experience of talking to an AI that feels like you're talking to a human that is super knowledgeable and super smart.

There's so much that's happened in public conversation around AI.

And there's so many questions that I have.

I've been speaking to a few people about artificial intelligence to try and understand it.

And I think what I am right now is I feel guite scared.

But when I get scared, I don't get it's not the type of scared that makes me anxious.

It's not like an emotional scared.

It's a very logical scared.

A very logical brain hasn't been able to figure out how the inevitable outcome that I've arrived at, which is that humans become the less dominant species on this planet.

How that is to be avoided in any way.

The first chapter of your book, The Coming Wave is titled appropriately to how I feel containment is not possible.

You say in that chapter, the widespread emotional reaction I was observing is something I've come to call the pessimism aversion trap.

Correct.

What is the pessimism aversion trap?

Well, so all of us be included.

Feel what you just described when you first get to grips with the idea of this new coming wave.

It's scary.

It's petrifying.

It's threatening.

Is it going to take my job?

Is my daughter or son going to fall in love with it?

You know, what does this mean?

What does it mean to be human in a world where there's these other human like things that aren't human?

How do I make sense of that?

It's super scary.

And a lot of people over the last few years, I think things have changed in the last six months, I have to say, but over the last few years, I would say the default reaction has been to avoid the pessimism. The pessimism and the fear, right, to just kind of recoil from it and pretend that it's like either not

happening or that it's all going to work out to be rosy.

It's going to be fine.

We don't have to worry about it.

People often say, well, we've always created new jobs.

We've never permanently displaced jobs.

We've only ever seen new jobs be created.

Unemployment is at an all time low.

Right.

So there's this default optimism bias that we have.

And I think it's less about a need for optimism and more about a fear of pessimism.

And so that trap, particularly in elite circles, means that often we aren't having the tough conversations that we need to have in order to respond to the coming wave.

Are you scared in part about having those tough conversations because of how it might be received? Not so much anymore.

So I've spent most of my career trying to put those tough questions on the policy table, right?

I've been raising these questions, the ethics of AI, safety and questions of containment for as long as I can remember with governments and civil societies and all the rest of it.

And so I've become used to talking about that.

And, you know, I think it's essential that we have the honest conversation because we can't let it happen to us.

We have to openly talk about it.

I mean, this is a big question, but as you sit here now, do you think that it is containable?

Because I can't see how.

I can't see how it can be contained.

Chapter three is the containment problem.

Or you give the example of how technologies are often invented for good reasons and for certain use cases, like the hammer, you know, which is used, you know, maybe to build something, but it also can be used to kill people.

And you say in history, we haven't been able to ban a technology ever.

Really, it has always found a way into society.

Because if other societies have an incentive to have it, even if we don't, and then we need we need it like the nuclear bomb, because if they have it, then we don't, then we're at disadvantage.

So are you optimistic?

Honestly,

I don't think an optimism or a pessimism frame is the right one, because both are equally biased in ways that I think distract us.

As I say in the book, on the face of it, it does look like containment isn't possible.

We haven't contained or permanently banned a technology of this type in the past.

There are some that we have done.

So we banned CFCs, for example, because they were producing a hole in the ozone layer.

We've banned certain weapons, chemical and biological weapons, for example, or blinding lasers, believe it or not.

There are such things as lasers that were instantly blind you.

So we have stepped back from the frontier in some cases, but that's largely where there's either cheaper or equally effective alternatives that are quickly adopted.

In this case, these technologies are omni-use.

So the same core technology can be used to identify cancerous tumors in chest X-rays as it can to identify a target on the battlefield for an aerial strike.

So that mixed use or omni-use is going to drive the proliferation, because there's huge commercial incentives, because it's going to deliver a huge benefit and do a lot of good.

And that's the challenge that we have to figure out is how to stop something which on the face of it is so good, but at the same time can be used in really bad ways too.

Do you think we will?

I do think we will.

So I think that nation states remain the backbone of our civilization.

We have chosen to concentrate power in a single authority, the nation state, and we pay our taxes and we've given the nation state a monopoly over the use of violence.

And now the nation state is going to have to update itself quickly to be able to contain this technology.

Because without that kind of essentially oversight, both of those of us who are making it, but also crucially of the open source, then it will proliferate and it will spread.

But regulation is still a real tool and we can use it and we must.

What does the world look like in, let's say, 30 years if that doesn't happen in your view?

Because the average person can't really grapple their head around artificial intelligence.

When they think of it, they think of these large language models that you can chat to and ask it about your homework.

That's the average person's understanding of artificial intelligence because that's all they've ever been exposed to of it.

You have a different view because of the work you spent the last decade doing.

So to try and give Dave, who's, I don't know, an Uber driver in Birmingham, an idea who's listening to this right now, what artificial intelligence is and its potential capabilities if there's no containment.

What does the world look like in 30 years?

So I think it's going to feel largely like another human.

So think about the things that you can do, not again in the physical world, but in the digital world.

2050, I'm thinking of. I'm in 2050.

2050, we will have robots.

2050, we will definitely have robots.

I mean, more than that 2050, we will have new biological beings as well.

Because the same trajectory that we've been on with hardware and software is also going to apply to the platform of biology.

Are you uncomfortable talking about this?

Yeah. I mean, it's pretty wild, right?

I know you should cross your arms.

No, I always look, I always use that as a cue for someone when a subject matter is uncomfortable.

And it's interesting because I know you know so much more than me about this.

And I know you've spent way more hours thinking off into the future about the consequences of this. I mean, you've written a book about it.

So you spent 10 years at the very deep mind is one of the pinnacle companies, the pioneers in this whole space.

So you know, you know some stuff.

And it's funny because when I was, I watched an interview with Elon Musk and he was asked a question similar to this.

I know he speaks in certain certain tone of voice, but he said that he's almost, he's gotten to the point where he thinks he's living in suspended disbelief.

Where he thinks that if he spent too long thinking about it, he wouldn't understand the purpose of what he's doing right now.

And he says that it's more dangerous than nuclear weapons and that it's too late too late to stop it. There's one interview that's chilling and I was filming dragons den the other day and I showed the dragons the clip.

I was like, look what Elon Musk said when he was asked about what his chart, what advice he should give to his children in a world of, in an inevitable world of artificial intelligence.

It's the first time I've seen Elon Musk stop for like 20 seconds and not know what to say, stumble, stumble, stumble, stumble, and then conclude that he's living in suspended disbelief. Yeah, I mean, I think it's a great phrase.

That is the moment we're in.

We have to, so what I said to you about the pessimism version trap, we have to confront the probability of seriously dark outcomes and we have to spend time really thinking about those consequences.

Because the competitive nature of companies and of nation states is going to mean that every organization is going to race to get their hands on intelligence.

Intelligence is going to be a new form of capital, right?

Just as there was a grab for land or there's a grab for oil, there's a grab for anything that enables you to do more with less, faster, better, smarter, right?

And we can clearly see the predictable trajectory of the exponential improvements in these technologies.

And so we should expect that wherever there is power, there's now a new tool to amplify that power, accelerate that power, turbocharge it, right?

And, you know, in 2050, if you ask me to look out there, I mean, of course it makes me grimace.

That's why I was like, oh my God, it's, it really does feel like a new species.

And that has to be brought under control.

We cannot allow ourselves to be dislodged from our position as the dominant species on this planet. We cannot allow that.

You mentioned robots.

So these are sort of adjacent technologies that are rising with artificial intelligence, robots.

You mentioned biological, new biological species.

Give me some light on what you mean by that.

Well, so so far the dream of robotics hasn't really come to fruition, right?

I mean, we still have, the most we have now are sort of drones and a little bit of self-driving cars.

But that is broadly on the same trajectory as these other technologies.

And I think that over the next 30 years, you know, we are going to have humanoid robotics.

We're going to have, you know, physical tools within our everyday system that we can rely on that will be pretty good,

that will be pretty good to do many of the physical tasks.

And that's a little bit further out because I think it, you know, there's a lot of tough problems there, but it's still coming in the same way.

And likewise with biology, you know, we can now sequence a genome for a millionth of the cost of the first genome, which took place in 2000.

So 20-ish years ago, the cost has come down by a million times.

And we can now increasingly synthesize that is create or manufacture new bits of DNA, which obviously give rise to life in every possible form.

And we're starting to engineer that DNA to either remove traits or capabilities that we don't like, or indeed to add new things that we want it to do.

We want, you know, fruit to last longer, or we want meat to have higher protein, et cetera, et cetera. Synthetic meat to have higher protein levels.

And what's the implications of that?

Well, potential implications.

I think that the darkest scenario there is that people will experiment with pathogens, engineered, you know, synthetic pathogens that might end up accidentally or intentionally being

I.e., they can spread faster or more lethal.

I.e., you know, they cause more harm or potentially kill.

Like a pandemic.

more transmissible.

Like a pandemic.

And that's where we need containment, right?

We have to limit access to the tools and the know-how to carry out that kind of experimentation. So one framework of thinking about this with respect to making containment possible is that we really are experimenting with dangerous materials.

And anthrax is not something that can be bought over the Internet, that can be freely experimented with.

And likewise, the very best of these tools in a few years' time are going to be capable of creating, you know, new synthetic pandemic pathogens.

And so we have to restrict access to those things.

That means restricting access to the compute.

It means restricting access to the software that runs the models, to the cloud environments that provide APIs,

provide you access to experiment with those things.

And of course, on the biology side, it means restricting access to some of the substances.

And people aren't going to like this.

People are not going to like that claim because it means that those who want to do good with those tools.

those who want to create a startup, the small guy, the little developer that struggles to comply with all the regulations,

they're going to be pissed off, understandably, right?

But that is the age we're in.

Deal with it.

Like, we have to confront that reality.

That means that we have to approach this with the precautionary principle, right?

Never before in the invention of a technology or in the creation of a regulation have we proactively said.

we need to go slowly.

We need to make sure that this first does no harm, the precautionary principle.

And that is just an unprecedented moment.

No other technology has done that, right?

Because I think we collectively in the industry, those of us who are closest to the work,

can see a place in five years or 10 years where it could get out of control.

And we have to get on top of it now.

And it's better to forego like that is give up some of those potential upsides or benefits until we can be more sure that it can be contained,

that it can be controlled, that it always serves our collective interests.

And I think about that.

So I think about what you've just said there about being able to create these pathogens, these diseases and viruses, etc.

That, you know, could become weapons or whatever else.

But with artificial intelligence and the power of that intelligence, with these pathogens,

you could theoretically ask one of these systems to create a virus that a very deadly virus.

You could ask the artificial intelligence to create a very deadly virus that has certain properties, maybe even that mutates over time in a certain way.

So it only kills a certain amount of people, kind of like a nuclear bomb of viruses that you could just pop, hit an enemy with.

Now, if I hear that and I go, OK, that's powerful, I would like one of those.

You know, there might be an adversary out there that goes, I would like one of those just in case America get out of hand.

And America is thinking, you know, I want one of those in case Russia gets out of hand.

And so, OK, you might take a precautionary approach in the United States,

but that's only going to put you on the back foot when China or Russia or one of your adversaries accelerates forward in that path.

And this was the same with the nuclear bomb.

And, you know, you nailed it.

I mean, that is the race condition.

We refer to that as the race condition, the idea that if I don't do it, the other party is going to do it. And therefore I must do it.

But the problem with that is that it creates a self-fulfilling prophecy.

So the default there is that we all end up doing it.

And that can't be right because there is a opportunity for massive cooperation here.

There's a shed that is between us and China and every other, quote unquote, them or they or enemy that we want to create.

We've all got a shared interest in advancing the collective health and well-being of humans and humanity.

How well have we done at promoting shared interest?

Well, in the development of technologies over the years, even at like a corporate level, even, you know.

You know, the nuclear non-proliferation treaty has been reasonably successful.

There's only nine nuclear states in the world today.

We've stopped, like three countries actually gave up nuclear weapons because we incentivize them with sanctions and threats and economic rewards.

Small groups have tried to get access to nuclear weapons and so far have largely failed.

It's expensive though, right?

And hard to, like uranium as a chemical to keep it stable and to buy it and to house it.

I mean, I couldn't just put it in the shed.

You certainly couldn't put it in a shed.

You can't download Uranium 235 off the internet.

It's not available open source.

That is totally true.

So it's got different characteristics for sure.

But a kid in Russia could, you know, in his bedroom could download something onto his computer.

That's incredibly harmful in the artificial intelligence department, right?

I think that that will be possible at some point in the next five years.

It's true because there's a weird trend that's going on here.

On the one hand, you've got the cutting edge AI models that are built by Google and open AI and my company inflection and they cost hundreds of millions of dollars.

And there's only a few of them.

But on the other hand, the what was cutting edge a few years ago is now open source today.

So GPT three, which came out in the summer of 2020 is now reproduced as an open source model.

So the code and the weights of the model, the design of the model and the actual implementation code is completely freely available on the web.

And it's tiny.

It's like 60 times or 60, 70 times smaller than the original model, which means that it's cheaper to use and cheaper to run.

And that's as you know, we've said earlier, like that's the natural trajectory of technologies that become useful.

They get more efficient, they get cheaper and they spread further.

And so that's the containment challenge.

That's really the essence of what I'm sort of trying to raise in my book is to frame the challenge of the next 30 to 50 years as around containment and around confronting proliferation.

Because we're both going to be alive unless this, you know, there's some robot kills us, but we're both going to be alive in 30 years time.

I hope so.

Maybe the podcast will still be going unless AI is now taking my job.

It's very possible.

So I'm going to sit you here and, you know, when you're 60, 68 years old, I'll be 60.

And I'll say at that point when we have that conversation, do you think we would have been successful in containment at a global level?

I think we have to be.

I can't even think that we're not.

Whv?

Why?

Because I'm fundamentally a humanist.

And I think that we have to make a choice to put our species first.

And I think that that's what we have to be defending for the next 50 years.

That's what we have to defend.

Because, look, it's, it's certainly possible that we invent these AGI's in such a way that they are always going to be provably subservient to humans and take instructions, you know, from their human controller every single time.

But enough of us think that we can't be sure about that, that I don't think we should take the gamble, basically.

So that's why I think that we should focus on containment and nonproliferation, because some people, if they do have access to the technology, will want to take those risks.

And they will just want to see, like, what's on the other side of the door, you know, and they might end up opening Pandora's box.

And that's a decision that affects all of us.

And that's the challenge of the networked age.

You know, we live in this globalized world and we use these words like globalization, and you sort of forget what globalization means.

This is what globalization is. This is what a networked world is.

It means that someone taking one small action can suddenly spread everywhere instantly.

Regardless of their intentions when they took the action.

It may be, you know, unintentional, like you say.

It may be that they never, they weren't ever meaning to do harm.

Well, I think I asked you when I said, you know, 30 years time, you said that there will be like human level intelligence.

You'll be interacting with, you know, this new species.

But the species, for me to think the species will want to interact with me is feels like wishful thinking.

Because what will I be to them?

You know, like, I've got a French bulldog Pablo.

And I can't imagine our IQ is that far apart.

Like, you know, in relative terms, the IQ between me and my dog Pablo, I can't imagine that's that far apart.

Even when I think about, is it like the orangutan where we only have like one percent difference in DNA or something crazy?

And yet they throw their poop around and I'm sat here broadcasting around the world.

There's guite a difference in that one percent, you know.

And then I think about this new species where as you write in your book in chapter four,

there seems to be no upper limit to AI's potential intelligence.

Why would such an intelligence want to interact with me?

Well, it depends how you design it.

So I think that our goal, one of the challenges of containment is to design

Als that we want to interact with that want to interact with us, right?

If you set an objective function for an AI, a goal for an AI by its design,

which, you know, inherently disregards or disrespects you as a human and your goals, then it's going to wander off and do a lot of strange things.

Whatever has kids and the kids.

You know what I mean?

What if it replicates in a way where because I've heard this this conversation around like it depends how we design it.

You know, I think about it's kind of like if I have a kid and the kid grows up to be a thousand times more intelligent than me.

To think that I can have any influence on it on it when it's a thinking sentient developing species again feels like I'm overestimating my version of intelligence and importance and significance in the face of something that is incomprehensibly

like even a hundred times more intelligent than me.

And the speed of its computation is a thousand times what my the meat in my skull can do. Yeah.

Like how is it going to how do I know it's going to respect me or care about me or understand, you know, that I may, you know.

I think that comes back down to the containment challenge.

I think that if we can't be confident that it's going to respect you and understand you and work for you and us as a species overall, then that's where we have to adopt the precautionary principle.

I don't think we should be taking those kinds of risks in experimentation and design.

And now I'm not saying it's possible to design an AI that doesn't have those self improvement capabilities in the limit in like 30 or 50 years.

I think, you know, that's kind of what I was saying is like it seems likely that if you have one like that, it's going to take advantage of infinite amounts of data and implements of computation and it's going to kind of outstrip our ability to act.

And so I think we have to step back from that precipice.

That's what the containment problem is, is that it's it's actually saying no.

Sometimes it's saying no, and that's a different sort of muscle that we've never really exercised as a civilization.

And that's obviously why containment appears not to be possible because we've never done it before. We've never done it before. And every inch of our, you know, commerce and politics and our war and all of our instincts are just like clash, compete, clash, compete.

Profit.

Profit.

Grow, beat.

Exactly.

Dominate, you know, fear them, be paranoid.

Like now all this nonsense about like China being this new evil, like how does that slip into our culture?

How are we suddenly all shifted from thinking it's the Muslim terrorists about to blow us all up to now it's the Chinese who are about to, you know, blow up Kansas.

It's just like, what are we talking about?

Like, we really have to pair back the paranoia and the fear and the othering because those are the

incentive dynamics that are going to drive us to, you know, cause self harm to humanity.

Thinking the worst of each other.

There's a couple of key moments when in my understanding of artificial intelligence that have been kind of paradigm paradigm shifts for me because I think like many people, I thought of artificial intelligence as, you know, like a, like a child I was raising and I would program.

I'd code it to do certain things.

So I'd code it to play chess and I would tell it the moves that are conducive with being successful in chess.

And then I remember watching that like AlphaGo documentary, which I think was deep, deep mind was us.

Yeah, you guys.

So you programmed this, this artificial intelligence to play the game go, which is kind of like, just think of it kind of like a chess or a blackout or whatever.

And it eventually just beats the best player in the world of all time.

And the way it learned how to beat the best player in the world of all time, the world champion, who was by the way depressed when he got beat was just by playing itself, right?

And then there's this moment, I think in, is it game four or something where it does this move that no one could have predicted as a move that seemingly makes absolutely no sense.

Right.

In those moments where no one trained it to do that and it did something unexpected beyond where humans are trying to figure it out in hindsight.

This is where I go, how do you, how do you train it?

If it's doing things we didn't anticipate.

Right.

Like how do you control it when it's doing things that humans couldn't anticipate it doing? Well, we're looking at that move.

It's called like move 37 or something.

Correct. Yeah.

Is it move 37?

It is.

Nice intelligence.

Nice work.

I'm going to survive a bit longer than I thought.

It's like move 37.

You've got at least another decade in you.

Move 37 does this crazy thing and you see everybody like lean in and go, why has it done that?

And it turns out to be brilliant that humans couldn't, couldn't have forecasted.

The commentator actually thought it was a mistake.

He was a pro and he was like, there's definitely a mistake.

The AlphaGoes lost the game.

But it was so far ahead of us that it knew something we didn't.

Right.

Right.

That's when I lost hope in this whole idea of like, oh, train it to do what we want.

Like a dog like sit, paw, roll over.

Right.

Well, the real challenge is that we actually want it to do those things.

When it discovers a new strategy or it invents a new idea or it helps us find like, you know, a cure for some disease.

Right.

That's why we're building it.

Right.

Because we're reaching the limits of what we as, you know, humans can invent and solve.

Right.

Especially with what we're facing of, you know, in terms of population growth over the next 30 years and how climate change is going to affect that and so on.

We really want these tools to turbocharge us.

Right.

And yet, like it's that creativity and that invention, which obviously makes us also feel well, maybe it is really going to do things that we don't like for sure.

Right.

So interesting.

How do you contend with all of this?

How do you contend with the clear upside and then you must like Elon must be completely aware of the horrifying existential risk at the same time.

And you're building a big company in this space, which I think is valued at four billion now, Inflection AI, which has got this, its own model called Pi.

So you're building in this space.

You understand the incentives at both a nation state level and a corporate level that we're going to keep planning forward.

Even if the US stops, there's going to be some other country that sees that as a huge advantage.

Their economy will swell because they did.

If this company stops, then this one's going to get a huge advantage in their shareholders.

You know, everyone's investing in AI full steam ahead, but you feel you can see this huge existential risk.

Is it suspended?

Is that the path forward suspended disbelief?

I mean, just to kind of like just know that it's like, I feel like I know that's going to happen.

No one's been able to tell me otherwise.

But just don't think too much about it and you'll be okay.

I think you can't give up.

Right.

I think that in some ways, your realization, exactly what you've just described, like weighing up to conflicting and horrible truths about what is likely to happen, those contradictions,

that is a kind of honesty and a wisdom.

I think that we need all collectively to realize.

Because the only path through this is to be straight up and embrace, you know, the risks and embrace the default trajectory of all these competing incentives driving forward to kind of make this feel like inevitable.

And if you put the blinkers on and you kind of just ignore it, or if you just be super rosy and it's all going to be all right.

And if you say that we've always figured it out anyway, then we're not going to get the energy and the dynamism and engagement from everybody to try to figure this out.

And that's what gives me like reason to be hopeful.

Because I think that we make progress by getting everybody paying attention to this. It isn't going to be about those who are currently the AI scientists or those who are the technologists,

you know, like me or the venture capitalists or just the politicians, like all of those people, no one's got answers.

So that's what we have to confront.

There are no obvious answers to this profound question.

And I've basically written the book to say, prove that I'm wrong.

You know, containment must be possible.

And I...

It must be.

It must be possible.

Why?

It has to be possible.

You want it to be.

I desperately want it to be.

Why must it be?

Because otherwise I think you're in the camp of believing that this is the inevitable evolution of humans, the transhuman kind of view.

You know, some people would argue like, what is...

Okay, let's stretch the timelines out.

So let's not talk about 30 years.

Let's talk about 200 years.

Like, what is this going to look like in 2200?

You tell me, you're smarter than me.

I mean, it's mind blowing.

It's mind blowing.

What is the answer?

We'll have quantum computers by then.

What's a quantum computer?

A quantum computer is a completely different type of computing architecture, which in simple terms basically allows you to...

Those calculations that I described at the beginning, billions and billions of flops,

those billions of flops can be done in a single computation.

So everything that you see in the digital world today relies on computers processing information.

And the speed of that processing is a friction.

It kind of slows things down, right?

You remember back in the day, old school modems, 56K modem, the dial-up sound and the image pixel loading, like pixel by pixel, that was because the computers were slow.

And we're getting to a point now where the computers are getting faster and faster and faster, and quantum computing is like a whole new leap, like way, way, way beyond where we currently are.

And so...

By analogy, how would I understand that?

I've got my dial-up modem over here, and then quantum computing over here.

How do I...

What's the difference?

Well. I don't know.

It's really difficult to explain.

Is it like a billion times faster?

Oh, it's like billions of billions of times faster.

It's much more than that.

I mean, one way of thinking about it is like a floppy disk, which I guess most people remember, 24 megabytes, a physical thing back in the day.

In 1960 or so, that was basically an entire pallet's worth of computer that was moved around by a forklift truck, right?

Which is insane.

Today, you know, you have billions and billions of times that floppy disk in your smartphone, in your pocket.

Tomorrow, you're going to have billions and billions of smartphones in miniscule wearable devices.

There'll be cheap fridge magnets that, you know, are constantly on everywhere, sensing all the time, monitoring, processing, analyzing, improving, optimizing, you know, and they'll be super cheap.

So, it's super unclear what do you do with all of that knowledge and information.

I mean, ultimately, knowledge creates value.

When you know the relationship between things, you can improve them, you know, make it more efficient.

And so, more data is what has enabled us to build all the value of, you know, online in the last 25 years.

And so, what does that look like in 150 years?

I can't really even imagine, to be honest with you.

It's very hard to say.

I don't think everybody is going to be working.

We wouldn't be working in that kind of environment.

I mean, the other trajectory to add to this is the cost of energy production.

You know, AI, if it really helps us solve battery storage, which is the missing piece, I think, to really tackle climate change, then we will be able to source, basically,

source and store infinite energy from the sun.

And I think in 20 or so years' time, 20, 30 years' time, that is going to be a cheap and widely available, if not completely freely available, resource.

And if you think about it, everything in life has the cost of energy built into its production value.

And so, if you strip that out, everything is likely to get a lot cheaper.

We'll be able to desalinate water.

We'll be able to grow crops much, much cheaper.

We were able to grow much higher quality food, right?

It's going to power new forms of transportation.

It's going to reduce the cost of drug production and healthcare, right?

So, all of those gains, obviously, there will be a huge commercial incentive to drive the production of those gains, but the cost of producing them is going to go through the floor.

I think that's one key thing that a lot of people don't realize.

There is a reason to be hugely hopeful and optimistic about the future.

Everything is going to get radically cheaper in 30 to 50 years.

So, 200 years' time, we have no idea what the world looks like.

This goes back to the point about being...

Did you say transhumanist?

Right.

What does that mean?

Transhumanism...

I mean, it's a group of people who basically believe that humans and our soul and our being will one day transcend or move beyond our biological substrate.

So, our physical body, our brain, our biology is just an enabler for your intelligence and who you are as a person.

And there's a group of kind of crackbots, basically, I think, who think that we're going to be able to upload ourselves to a silicon substrate, right?

A computer that can hold the essence of what it means to be Steven.

Why are they wrong?

I think about all these adjacent technologies like biological advancements.

Did you call it like biosynthesis or something?

Yeah, synthetic biology.

Synthetic biology.

I think about the nanotechnology development.

Right.

I think about quantum computing.

The progress in artificial intelligence, everything becoming cheaper.

And I think, why are they wrong?

It's hard to say precisely, but broadly speaking, I haven't seen any evidence yet that we're

able to extract the essence of a being from a brain, right?

It's that kind of dualism that, you know, we're going to be able to extract the essence of that kind of dualism that, you know, there is a mind and a body and a spirit.

I don't see much evidence for that, even in neuroscience.

But actually, it's much more one and the same.

So I don't think, you know, you're going to be able to emulate the entire brain.

So their thesis is that, well, some of them cryogenically store their brain after death. Iesus.

They wear these like, you know, how you have like an organ donor tag or whatever.

So they have a cryogenically freeze me when I die tag.

And so there's like a special ambulance services that will come pick you up.

Because obviously you need to do it really quickly.

The moment you die, you need to get put into a cryogenic freezer to preserve your brain forever.

I personally think this is nuts.

But, you know, their belief is that you'll then be able to reboot that biological brain and transfer you over.

It doesn't seem plausible to me.

When you said at the start of this, this little topic here that you,

it must be possible to contain it.

It must be possible.

The reason why I struggle with that is because in chapter seven,

you say a line in your book that AI is more autonomous than any other technology in history.

For centuries, the idea that technologies is somehow running out of control,

a self-directed and self-propelling force beyond the realms of human agency

remained a fiction, not anymore.

And this idea of autonomous technology that is acting uninstructed and is intelligent.

And then you say, we must be able to contain it.

It's kind of like a massive dog, like a big rottweiler.

That is, you know, a thousand times bigger than me.

And me looking up at it and going, I'm going to take you for a walk.

And then it's just looking down at me and just stepping over me, or stepping on me.

Well, that's actually a good example because we have actually contained rottweilers before.

We've contained gorillas and, you know, tigers and crocodiles and pandemic pathogens and nuclear weapons.

And so, you know, it's easy to be, you know, a hater on what we've achieved.

But this is the most peaceful moment in the history of our species.

This is a moment when our biggest problem is that people eat too much.

Think about that.

We've spent our entire evolutionary period running around looking for food

and trying to stop, you know, our enemies throwing rocks at us.

And we've had this incredible period of 500 years where, you know, each year things have broadly, maybe each century, let's say, there's been a few ups and downs, but things have broadly got better.

And we're on a trajectory for, you know, life spans to increase and quality of life to increase and health and well-being to improve.

And I think that's because in many ways we have succeeded in containing forces that appear to be more powerful than ourselves.

It just requires unbelievable creativity and adaptation.

It requires compromise and it requires a new tone, right?

A much more humble tone to governance and politics and how we run our world.

Not this kind of like hyper-aggressive, adversarial, paranoia tone that we talked about previously, but one that is like much more wise than that, much more accepting that we are unleashing this force

that does have that potential to be the Rottweiler that you described,

but that we must contain that as our number one priority.

That has to be the thing that we focus on because otherwise it contains us.

I've been thinking a lot recently about cybersecurity as well, just broadly on an individual level.

In a world where there are these kinds of tools, which seems to be quite close, large language models.

it brings up this whole new question about cybersecurity and cybersecurity and, you know,

in a world where there's the ability to generate audio and language and videos that seem to be real, what can we trust?

And, you know, I was watching a video of a young girl whose grandmother was called up by a voice that was made to sound like her son saying he'd been in a car accident and asking for money and her nearly sending the money.

Or this whole, you know, because this really brings into focus that we, our lives are built on trust, trusting the things we see here and watch.

And in now, it feels like a moment where we're no longer going to be able to trust what we see on the internet, on the phone.

What advice do you have for people who are worried about this?

So skepticism, I think, is healthy and necessary.

And I think that we're going to need it even more than we ever did, right?

And so if you think about how we've adapted to the first wave of this, which was spammy email scams,

everybody got them, and over time, people learned to identify them and be skeptical of them and reject them.

Likewise, you know, I'm sure many of us get, like, text messages.

I'd certainly get loads of text messages trying to fish me and ask me to meet up or do this, that and the other.

And we've adapted, right?

Now, I think we should all know and expect that criminals will use these tools to manipulate us, just as you described.

I mean, you know, the voice is going to be human-like.

The deep fake is going to be super convincing.

And there are actually ways around those things.

So, for example, the reason why the banks invented one-time passwords, where they send you a text

message with a special code,

is precisely for this reason, so that you have a two FA, a two-factor authentication.

Increasingly, we will have a three or four-factor authentication, where you have to triangulate between multiple separate independent sources.

And it won't just be, like, call your bank manager and release the funds, right?

So, this is where we need the creativity and energy and attention of everybody, because defence, the kind of defensive measures,

have to evolve as quickly as the potential offensive measures, the attacks that are coming.

I heard you say this, that you think some people, for many of these problems,

were going to need to develop AIs to defend us from the AIs.

Right. We kind of already have that, right?

So, we have automated ways of detecting spam online these days.

You know, most of the time, there are machine learning systems which are trying to identify when your credit card is used in a fraudulent way.

That's not a human sitting there looking at patterns of spending traffic in real time.

That's an AI that is, like, flagging that something looks off.

Likewise, with data centres or security cameras, a lot of those security cameras these days are, you know,

have tracking algorithms that look for, you know, surprising sounds, or, like, if a glass window is smashed,

that will be detected by an AI often that is, you know, listening on the security camera.

So, you know, that's kind of what I mean by that, is that increasingly those AIs will get more capable, and we'll want to use them for defensive purposes.

And that's exactly what it looks like to have good, healthy, well-functioning, controlled AIs that serve us.

I went on one of these large language models and said to me, I said to the large language model, give me an example where artificial intelligence takes over the world or whatever and results in the destruction of humanity.

And then tell me what we'd need to do to prevent it.

And it said, it gave me this wonderful example of this AI called Cynthia that threatens to destroy the world.

And it says, the way to defend that would be a different AI, which had a different name.

And it said that this one would be acting in human interests and we'd basically be fighting one AI with another AI.

And of course, of course, at that level, if Cynthia started to wreak havoc on the world and take control of the nuclear weapons and infrastructure and all that, we would need an equally intelligent weapon to fight it.

Although one of the interesting things that we found over the last few decades is that it so far tended to be the AI plus the human that is still dominating.

That's the case in chess, in Go and other games.

Yeah, so there was a paper that came out a few months ago, two months ago, that showed that a human was actually able to beat the cutting edge Go program, even one that was better than AlphaGo with a new strategy that they had discovered.

So obviously, it's not just a sort of game over environment where the AI just arrives and it gets better.

Like humans also adapt, they get super smart, they, like I say, get more skeptical, ask good questions,

invent their own things, use their own AIs to adapt.

And that's the evolutionary nature of what it means to have a technology.

I mean, everything is a technology, like your pair of glasses made you smarter in a way.

Before there were glasses and people got bad eyesight, they weren't able to read.

Suddenly, those who did adopt those technologies were able to read for longer in their lives or under low light conditions and they were able to consume more information and got smarter.

And so that is the trajectory of technology.

It's this iterative interplay between human and machine that makes us better over time.

You know the potential consequences if we don't reach a point of containment,

yet you chose to build a company in this space.

Yeah.

Why that? Why did you do that?

Because I believe that the best way to demonstrate how to build safe and contained AI is to actually experiment with it in practice.

And I think that if we are just skeptics or critics and we stand back from the cutting edge, then we give up that opportunity to shape outcomes to all of those other actors that we referred to, whether it's like China or in the US going at each other's throats,

or other big companies that are purely pursuing profit at all costs.

And so it doesn't solve all the problems.

Of course, it's super hard and again, it's full of contradictions.

But I honestly think it's the right way for everybody to proceed.

Experiment at the front.

Yeah. If you're afraid...

China, Russia, Putin.

Understand, right?

What reduces fear is deep understanding.

Spend time playing with these models.

Look at their weaknesses.

They're not superhumans yet.

They make tons of mistakes.

They're crappy in lots of ways.

They're actually not that hard to make.

The more you've experimented, has that correlated with a reduction in fear?

Cheeky question.

Yes, and no, you're totally right.

Yes, it has in the sense that, you know, the problem is the more you learn,

the more you realize...

Yeah, that's what I'm saying.

I was fine before I started talking about AI.

Now, then why have talked about it?

It's true. It's true.

It's sort of pulling on a thread.

This is a crazy spiral.

Yeah, I mean, like, I think in the short term, it's made me way less afraid

because I don't see that kind of existential harm that we've been talking about

in the next decade or two.

But longer term, that's where I struggle to wrap my head around how things play out in 30 years.

Some people say government regulation will sort it out.

You discussed this in chapter 13 of your book where you, which is titled

Containment Must Be Possible.

I love how you didn't say is.

Containment must be possible.

What do you say to people that say government regulation will sort it out?

I had Rishi Sunak did some announcement and he's got a Cobra committee coming together.

They'll handle it.

That's right.

And the EU have a huge piece of regulation called the EU AI Act.

President Joe Biden has gotten his own set of proposals and we've been working with both Rishi Sunak and Biden

and trying to contribute and shape it in the best way that we can.

Look, it isn't going to happen without regulation.

So regulation is essential.

It's critical.

Again, going back to the precautionary principle.

But at the same time, regulation isn't enough.

You know, I often hear people say, well, we'll just regulate it.

We'll just stop.

We'll just stop.

We'll just stop.

We'll slow down.

And the problem with that is that it kind of ignores the fact that the people who are putting together the regulation

don't really understand enough about the detail today.

You know, in their defense, they're rapidly trying to wrap their head around it,

especially in the last six months.

And that's a great relief to me because I feel the burden is now increasingly shared.

And, you know, just from a personal perspective, I'm like, I feel like I've been saying this for about a decade.

And just in the last six months, now everyone's coming at me and saying, like, you know, what's going on?

I'm like, great.

This is the conversation we need to be having because everybody can start to see the glimmers of

the future.

Like, what will happen if a chat GPT-like product or a pie-like product really does improve over the next 10 years?

And so when I say, you know, regulation is not enough, what I mean is it needs movements.

It needs culture.

It needs people who are actually building and making, you know, in like modern, creative, critical ways,

not just like giving it up to, you know, companies or small groups of people, right?

We need lots of different people experimenting with strategies for containment.

Isn't it predicted that this industry is a \$15 trillion industry or something like that?

Yeah, I've heard that.

It is a lot.

So if I'm Rishi and I know that I'm going to be chucked out of office, Rishi's the Prime Minister of the UK.

If I'm going to be chucked out of office in two years, unless this economy gets good,

I don't want to do anything to slow down that \$15 trillion bag that I could be on the receiving end of.

I would definitely not want to slow that \$15 trillion bag and give it to like America or Canada or some other country.

I'd want that \$15 trillion windfall to be on my country.

Right.

So I have no other than the long-term, you know, health and success of humanity.

In my four-year election window, I've got to do everything I can to boost these numbers and get us looking good.

So I could give you lip service, but listen, I'm not going to be here unless these numbers look good. Right.

Exactly.

That's another one of the problems.

Short-termism is everywhere.

Who is responsible for thinking about the 20-year future?

Who is it?

I mean, that's a deep guestion, right?

I mean, the world is happening to us on a decade-by-decade timescale.

It's also happening hour by hour.

So change is just ripping through us.

And this arbitrary window of governance of like a four-year election cycle where actually it's not even four years,

because by the time you've got in, you do some stuff for six months.

And then by month, you know, 12 or 18, you're starting to think about the next cycle and are you going to pull, you know,

this is like the short-termism is killing us, right?

And we don't have an institutional body whose responsibility is stability.

You could think of it as like a, you know, like a global technology stability function.

What is the global strategy for containment that has the ability to introduce friction when necessary

to implement the precautionary principle

and to basically keep the peace?

That, I think, is the missing governance piece which we have to invent in the next 20 years.

And it's insane because I'm basically describing the UN Security Council,

plus the World Trade Organization, all these huge, you know, global institutions which formed after, you know, the horrors of the Second World War

have actually been incredible.

They've created interdependence and alignment and stability, right?

Obviously, there's been a lot of bumps along the way in the last 70 years, but broadly speaking, it's an unprecedented period of peace.

And when there's peace, we can create prosperity.

And that's actually what we're lacking at the moment, is that we don't have an international mechanism for coordinating among competing nations, competing corporations to drive the peace.

In fact, we're actually going kind of in the opposite direction.

We're resorting to the old-school language of a clash of civilizations with like, China is the new enemy.

They're going to come to dominate us.

We have to dominate them.

It's a battle between two poles.

China's taking over Africa.

China's taking over the Middle East.

We have to count.

I mean, it's just like that can only lead to conflict.

That just assumes that conflict is inevitable.

And so when I say regulation is not enough, no amount of good regulation in the UK or in Europe or in the U.S.

is going to deal with that clash of civilizations language, which we seem to have become addicted to. If we need that global collaboration to be successful here, are you optimistic now that we'll get it? Because the same incentives are at play with climate change and AI.

Why would I want to reduce my carbon emissions when it's making me loads of money?

Why would I want to reduce my AI development when it's going to make us 15 trillion?

So the really painful answer to that question is that we've only really ever driven extreme compromise and consensus in two scenarios.

One, off the back of unimaginable catastrophe and suffering, you know, Hiroshima and Nagasaki and the Holocaust and World War II,

which drove 10 years of consensus and new political structures, right?

And then the second is...

We did fire the bullet, though, didn't we?

We fired a couple of those nuclear bombs.

Exactly.

And that's why I'm saying the brutal truth of that is that it takes a catastrophe to trigger the need for alignment, right?

So that's one.

The second is where there is an obvious mutually assured destruction, you know, dynamic, where both parties are afraid that this would trigger nuclear meltdown, right?

And that means suicide.

And when there was few parties.

Exactly.

When there was just nine people.

Exactly.

You could get all nine.

But when we're talking about artificial technology, there's going to be more than nine people, right? That have power access to the false sort of power of that technology for nefarious reasons.

I don't think it has to be like that.

I think that's the challenge of containment is to reduce the number of actors that have access to the existential threat technologies to an absolute minimum and then use the existing military and economic incentives which have driven world order and peace so far

to prevent the proliferation of access to these super intelligences or these AGI's.

A quick word on Huell.

As you know, they're a sponsor of this podcast and I'm an investor in the company.

And I have to say it's moments like this in my life where I'm extremely busy and I'm flying all over the place and I'm recording TV shows and I'm recording shows in America and here in the UK that Huell is a necessity in my life.

I'm someone that regardless of external circumstances or professional demands wants to stay healthy and nutritionally complete and that's exactly where Huell fits in my life.

It's enabled me to get all of the vitamins and minerals and nutrients that I needed in my diet to be aligned with my health goals while also not dropping the ball on my professional goals because it's convenient and because I can get it online in Tesco in supermarkets all over the country.

If you're one of those people that hasn't yet tried Huell or you have before but for whatever reason, you're not a Huell consumer right now.

I would highly recommend giving Huell a go.

And Tesco have now increased their listings with Huell so you can now get the RTD ready to drink in Tesco Express all across the UK.

Ten areas of focus for containment.

You're the first person I've met that's really hazarded a laid out blueprint for the things that need to be done.

Cohesively to try and reach this point of containment.

So I'm super excited to talk to you about these.

The first one is about safety and you mentioned that's kind of what we talked about a little bit about there being AIs that are currently being developed to help contain other AIs.

Two audits which is being able to from what I understand being able to audit what's being built in these open source models.

Three choke points.

What's that?

Yeah, so choke points refers to points in the supply chain where you can throttle who has access to

what.

So on the internet today, everyone thinks of the internet as an idea this kind of abstract cloud thing that hovers around above our heads.

But really the internet is a bunch of cables.

Those cables are physical things that transmit information under the sea and those points, the endpoints can be stopped and you can monitor traffic.

You can control basically what traffic moves back and forth.

And then the second choke point is access to chips.

So the GPUs, graphics processing units, which are used to train these super large clusters.

I mean, we now have the second largest supercomputer in the world today.

At least, you know, just for this next six months, we will other people catch up soon, but we're ahead of the curve.

We're very lucky cost a billion dollars.

And those chips are really the raw commodity that we use to build these large language models.

And access to those chips is something that governments can, should and are, you know, restricting. That's a choke point.

You spent a billion dollars on a computer.

We did. Yeah.

A bit more than that, actually.

About 1.3.

A couple of years time, that'll be the price of an iPhone.

That's the problem. Everyone's going to have it.

Number six is quite curious.

You say that the need for governments to put increased taxation on AI companies to be able to find fund the massive changes in society, such as paying for reskilling and education.

Yeah.

You put massive tax on over here.

I'm going to go over here.

If you tax it, if I'm an AI company and you're taxing me heavily over here, I'm going to Dubai.

Yep.

Portugal.

Yep.

So if it's that much of a competitive disadvantage, I will not build my company where the taxation is high.

Right.

Right.

So the way to think about this is what are the strategies for containment?

If we're agreed that long term, we want to contain that is closed down, slow down control, both the proliferation of these technologies and the way the really big AIs are used.

Then the way to do that is to tax things, tax things, taxing things slows them down.

And that's what you're looking for provided you can coordinate internationally.

So you're totally right that, you know, some people will move to Singapore or to Abu Dhabi or Dubai or whatever.

The reality is that at least for the next, you know, sort of period, I would say 10 years or so, the concentrations of intellectual, you know, horsepower will remain the big mega cities. Right.

You know, I moved from London in 2020 to go to Silicon Valley and I started my new company in Silicon Valley because the concentration of talent there is overwhelming.

All the very best people are there in AI and software engineering.

So I think it's quite likely that that's going to remain the case for the foreseeable future.

But in the long term, you're totally right.

It's another coordination problem.

How do we get nation states to collectively agree that we want to try and contain, that we want to slow down?

Because as we've discussed with the proliferation of dangerous materials or on the military side, there's no use one person doing it or one country doing it.

If others race ahead and that's the conundrum that we face.

Iam

I don't consider myself to be a pessimist in my life.

I consider myself to be an optimist generally.

I think and I always, I think that as you've said, I think we have no choice but to be optimistic and I have faith in humanity.

We've done so much, so many incredible things and overcome so many things.

And I also think I'm really logical as in I'm the type of person that needs evidence to change my beliefs either way.

When I look at all of the whole picture, having spoken to you and several others on this subject matter, I see more reasons why we won't be able to contain the reasons why we will, especially when I dig into those incentives.

You talk about incentives at length in your book at different points and it's clear that all the incentives are pushing towards a lack of containment,

especially in the short and midterm, which tends to happen with new technologies.

In the short and midterm, it's like a land grab.

The gold is in the stream.

We all rush to get the shovels and the, you know, the sieves and stuff.

And then we realize the unintended consequences of that.

Hopefully not before it's too late.

In chapter eight, you talk about unstoppable incentives at play here.

The coming wave represents the greatest economic prize in history.

And scientists and technologists are all too human.

They crave status, success and legacy.

And they want to be recognized as the first and the best, their competitive and clever with a carefully nurtured sense of their place in the world and in history.

Right.

I look at you.

I look at people like Sam from OpenAI, Elon.

You're all humans with the same understanding of your place in history and status and success.

You don't want that, right?

Right.

There's a lot of people that maybe aren't is don't have as good a track record of you at doing the right thing, which you certainly have.

That will just want the status and the success and the money.

Incredibly strong incentives.

I always think about incentives as being the thing that you look at.

You want to understand how people behave.

All of the incentives on a geopolitical level, like on a global level, suggest that containment won't happen.

Am I right in that assumption?

All the incentives suggest containment won't happen in the short or midterm until there is a tragic event that makes us, forces us towards that idea of containment.

Or if there is a threat of mutually assured destruction.

Right.

So that, and that's the case that I'm trying to make is that let's not wait for something catastrophic to happen.

So it's self-evident that we all have to work towards containment.

Right.

I mean, you would have thought that the potential threat, the potential idea that COVID-19 was a side effect, let's call it, of a laboratory in Wuhan

that was exploring gain of function research, where it was deliberately trying to basically make the pathogen more transmissible.

You would have thought that warning to all of us, let's not even debate whether it was or wasn't, but just the fact that it's conceivable that it could be,

that should really, in my opinion, have forced all of us to instantly agree that this kind of research should just be shut down.

We should just not be doing gain of function research.

On what planet could we possibly persuade ourselves that we can overcome the containment problem in biology?

Because we've proven that we can't.

Because it could have potentially got out.

And there's a number of other examples of where it did get out of other diseases like foot and mouth disease back in the 90s in the UK.

So that didn't change our behavior.

Right.

Well, foot and mouth disease clearly didn't cause enough harm because it only killed a bunch of cattle.

And the COVID-19 pandemic, we can't seem to agree that it really was from a lab and not from a bunch of bats.

And so that's where I struggle where, you know, now you catch me in a moment where I feel angry and sad and pessimistic,

because to me, that's like a straightforwardly obvious conclusion that, you know, this is a type of

research that we should be closing down.

And I think we should be using these moments to give us insight and wisdom about how we handle other technology trajectories in the next few decades.

Should.

We should.

We should.

That's what I'm advocating for.

Must.

That's the best I can do.

I want to know will.

Will, I think the odds are low.

I can only do my best.

I'm doing my best to advocate for it.

I mean, you know, like I'll give you an example.

Like I think autonomy is a type of AI capability that we should not be pursuing.

Really?

Like autonomous cars and stuff?

Autonomous cars, I think are slightly different because autonomous cars operate within a much more constrained physical domain, right?

Like, you know, you really can.

The containment strategies for autonomous cars are actually quite reassuring.

Okay.

They have, you know, GPS control, you know, we know exactly all the telemetry and how exactly all of those, you know, components on board a car operate.

And we can observe repeatedly that it behaves exactly as intended.

Right.

Whereas I think with, with other forms of autonomy that people might be pursuing, like online, you know, where you have an AI that is like designed to self improve without any human oversight or a battlefield weapon, which, you know, like unlike a car hasn't been, you know, over that particular moment in the

battlefield millions of times, but is actually facing a new enemy every time, you know, every single time.

And we're just going to go and, you know, allow these autonomous weapons to have, you know, these autonomous military robots to have lethal force.

I think that's something that we should really resist.

I don't think we want to have autonomous robots that have lethal force.

You're a super smart guy.

And I, I struggle to believe that you're, you, you, because you demonstrate such a clear understanding of the incentives in your book that I struggle to believe that you don't think the incentives will win out, especially in the short near term.

And then the problem is in the short near term, as is the case with most of these waves is we, we wake up in 10 years time ago, how the hell did we get here? Right.

And why like, and we, and as you say, this precautionary approach of we should have rang the bell earlier, we should have sounded the alarm earlier, but we waltzed in with optimism.

Right.

And with that kind of aversion to confronting the realities of it.

And then we woke up in 30 years and we're on a leash.

Right.

And there's a big rottweiler and we're, we've lost control.

We've lost, you know, I, I, I would love to know someone as smart as you.

I don't, I don't believe can be, can believe that containment is possible.

And that's me just being completely honest.

I'm not saying you're lying to me, but I just can't see how someone as smart as you and in the know as you can believe that containment is going to happen.

Well, I didn't say it is possible.

I said it must be right, which is what we keep discussing.

That's an important distinction.

You know, on the face of it, look what I care about.

I care about science.

I care about facts.

I care about describing the world as I see it.

And what I've set out to do in the book is describe a set of interlocking incentives which drive a technology production process, which produces potentially really dangerous outcomes.

And what I'm trying to do is frame those outcomes in the context of the containment problem and say, this is the big challenge of the 21st century.

Containment is the challenge.

And if it isn't possible, then we have serious issues.

And on the face of it, like I've said in the book, I mean, the first chapter is called containment is not possible, right?

The last chapter is called containment must be possible for all our sakes.

It must be possible.

But that, but I agree with you that I'm not, I'm not saying it is.

I'm saying this is what we have to be working on.

We have no choice.

We have no choice but to work on this problem.

This is a critical problem.

How much of your time are you focusing on this problem?

Basically all my time.

I mean, building and creating is about understanding how these models work, what their limitations are, how to build it safely and ethically.

I mean, we have designed the structure of the company to focus on the safety and ethics aspects.

So for example, we are a public benefit corporation, right?

Which is a new type of corporation, which gives us a legal obligation to balance profit making with the consequences of our actions as a company on the rest of the world.

The way that we affect the environment, you know, the way that we affect people, the way that we

affect users that people who aren't users of our products.

And that's a really interesting, I think, an important new direction.

It's a new evolution in corporate structure because it says we have a responsibility to proactively do our best to do the right thing, right?

And I think that if you were a tobacco company back in the day or an oil company back in the day and your legal charter said that your directors are liable

if they don't meet the criteria of stewarding your work in a way that doesn't just optimize profit, which is what all companies are incentivized to do at the moment, talking about incentives.

But actually in equal measure, it tends to the importance of doing good in the world.

To me, that's an incremental but important innovation in how we organize society and how we incentivize our work.

So it doesn't solve everything. It's not a panacea, but that's my effort to try and take a small step in the right direction.

Do you ever get sad about it? About what's happening?

Yeah, for sure. For sure. It's intense.

It's intense. It's a lot to take in. It's a very real reality.

Does that weigh on you?

Yeah, it does. I mean, every day, every day. I mean, I've been working on this for many years now and it's, you know, it's emotionally a lot to take in.

It's hard to think about the far out future and how your actions today, our actions collectively, our weaknesses, our failures, that, you know, that irritation that I have that we can't learn the lessons from the pandemic, right? Like all of those moments where you feel the frustration of governments not working properly or corporations not listening or some of the obsessions that we have in culture where we're debating like small things, you know,

and you're just like, whoa, we need to focus on the big picture here.

You must feel a certain sense of responsibility as well that most people won't carry because you've spent so much of your life at the very cutting edge of this technology and you understand it better than most.

You can speak to it better than most so you have a greater chance than many at steering. That's a responsibility.

Yeah, I embrace that. I try to treat that as a privilege. I feel lucky to have the opportunity to try and do that.

There's this wonderful thing in my favorite theatrical play called Hamilton, where he says, history has its eyes on you. Do you feel that?

Yeah, I feel that. I feel that. I feel that. It's a good way of putting it. I do feel that.

You're happy, right?

Well, what is happiness?

What's the range of emotions that you can tend with on a frequent basis, if you're being honest? I think it's kind of exhausting and exhilarating in equal measure because, for me, it is beautiful to see people interact with AIs and get huge benefit out of it.

Every day now, millions of people have a super smart tool in their pocket that is making them wiser and healthier and happier, providing emotional support, answering questions of every type, making you more intelligent.

And so on the face of it, in the short term, that feels incredible. It's amazing what we're all building. But in the longer term, it is exhausting to keep making this argument and have been doing it for a long time.

And in a weird way, I feel a bit of a sense of relief in the last six months because after chat GBT and this wave feels like it started to arrive and everybody gets it.

So I feel like it's a shared problem now. And that feels nice.

And it's not just bouncing around in your head.

A little bit. It's not just in my head and a few other people at DeepMind and OpenAI and other places that have been talking about it for a long time.

Ultimately, human beings may no longer be the primary planetary drivers as we have become accustomed to being.

We are going to live in an epoch where the majority of our daily interactions are not with other people, but with AIs.

Page 284 of your book.

The last page.

Think about how much of your day you spend looking at a screen.

12 hours.

Pretty much, right? Whether it's a phone or an iPad or a desktop versus how much time you spend looking into the eyes of your friends and your loved ones.

And so to me, it's like we're already there in a way.

You know, what I meant by that was, you know, this is a world that we're kind of already in.

You know, the last three years, people have been talking about metaverse, metaverse, metaverse.

And the mischaracterization of the metaverse was that it's over there.

It was this like virtual world that we would all bop around in and talk to each other as these little characters.

But that was totally wrong. That was a complete misframing.

The metaverse is already here.

It's the digital space that exists in parallel time to our everyday life.

It's the conversation that you will have on Twitter or, you know, the video that you'll post on YouTube or this podcast that will go out and connect with other people.

It's that meta space of interaction, you know, and I use meta to mean beyond this space, not just that weird other over there space that people seem to point to.

And that's really what is emerging here.

It's this parallel digital space that is going to live alongside with and in relation to our physical world.

Your kids come to you. You got kids?

No, I don't have kids.

Your future kids, if you ever have kids, a young child walks up to you and asks that question that Elon was asked.

What should I do about with my future?

What should I pursue in the light of everything you know about how artificial intelligence is going to change the world and computational power and all of these things?

What should I dedicate my life to? What do you say?

I would say knowledge is power. Embrace, understand, grapple with the consequences. Don't look the other way when it feels scary and do everything you can to understand and participate and shape because it is coming.

And if someone's listening to this and they want to do something to help this battle for which I think you presented as a solution containment, what can the individual do?

Read, listen, use the tools, try to make the tools, understand the current state of regulation, see which organizations are organizing around it like, you know, campaign groups, activism groups, find solidarity, connect with other people, spend time online, ask these questions, mention it at the pub, ask your parents, ask your mum how she's reacting to talking to Alexa or whatever it is that she might do.

Pay attention. I think that's already enough.

And there's no need to be more prescriptive than that because I think people are creative and independent and it will be obvious to you what you as an individual feel you need to contribute in this moment.

Provided you're paying attention.

Last question. What if we fail and what if we succeed? What if we fail in containment and what if we succeed in containment of artificial intelligence?

I honestly think that if we succeed, this is going to be the most productive and the most meritocratic moment in the history of our species.

We are about to make intelligence widely available to hundreds of millions, if not billions of people, and that is all going to make us smarter and much more creative and much more productive.

And I think over the next few decades, we will solve many of our biggest social challenges. I really believe that. I really believe we're going to reduce the cost of energy production, storage and distribution to zero marginal cost.

We're going to reduce the cost of producing healthy food and make that widely available to everybody.

And I think the same trajectory with healthcare, with transportation, with education, I think that ends up producing radical abundance over a 30 year period.

And in a world of radical abundance, what do I do with my day?

I think that's another profound question and believe me, that is a good problem to have. If we can, absolutely.

We don't need meaning and purpose.

Oh man, that is a better problem to have than what we've just been talking about for the last like 90 minutes.

And I think that's wonderful. Isn't that amazing?

I don't know. The reason I'm unsure is because everything that seems wonderful has an unintended consequence.

I'm sure it does. We live in a world of food abundance in the West and our biggest problem is obesity.

So I'll take that problem in the grand scheme of everything.

Humans not need struggle. Do we not need that kind of meaningful voluntary struggle?

I think we'll create other opportunities to quest.

I think that's an easier problem to solve and I think it's an amazing problem.

Like many people really don't want to work, right?

They want to pursue their passion and their hobby and all the things that you talk about and so on. Absolutely. We're now, I think, going to be heading towards a world where we can liberate people from the shackles of work unless you really want to.

Universal basic income?

I've long been an advocate of UBI. Very long time.

Everyone gets a check every month?

I don't think it's going to guite take that form.

I actually think it's going to be that we basically reduce the cost of producing basic goods so that you're not as dependent on income.

Imagine if you did have basically free energy and food.

You could use that free energy to grow your own food. You could grow it in a desert because you would have adapted seeds and so on.

You would have desalination and so on.

That really changes the structure of cities. It changes the structure of nations.

It means that you really can live in quite different ways for very extended periods without contact with the kind of center.

I'm actually not a huge advocate of that kind of libertarian wet dream, but I think if you think about it in theory, it's kind of a really interesting dynamic.

That's what proliferation of power means.

Power isn't just about access to intelligence. It's about access to these tools which allow you to take control of your own destiny and your life and create meaning and purpose in the way that you might envision.

And that's incredibly creative, incredibly creative time. That's what success looks like to me.

In some ways the downside of that, I think failure is not achieving a world of radical abundance in my opinion.

More importantly, failure is a failure to contain.

What does that lead to?

I think it leads to a mass proliferation of power and people who have really bad intentions will potentially use that power to cause harm to others.

This is part of the challenge, right?

In this networked, globalized world, a tiny group of people who wish to deliberately cause harm are going to have access to tools that can instantly, quickly have large-scale impact on many, many other people.

And that's the challenge of proliferation, is preventing those bad actors from getting access to the means to completely destabilize our world.

That's what containment is about.

We have a closing tradition on this podcast where the last guest leaves a question for the next guest, not knowing who they're leaving the question for.

The question left for you is, what is a space or place that you consider the most sacred?

Well, I think one of the most beautiful places I remember going to as a child was Windermere Lake in the Lake District.

And I was pretty young and on a dingy with some family members, and I just remember it being

incredibly serene and beautiful and calm.

I actually haven't been back there since, but that was a pretty beautiful place.

Seems like the antithesis of the world we live in.

Right. Maybe I should go back there and chill out.

Maybe. Thank you so much for writing such a great book.

It's wonderful to read a book on this subject matter that does present solutions, because not many of them do.

And it presents them in a balanced way that appreciates both sides of the argument.

Doesn't, isn't tempted to just play to either, what do they call it, playing to like the crowd?

No, they call it like playing to the orchestra, can't remember.

But just, it doesn't attempt to play to either side or ponder to either side in order to score points.

It seems to be entirely nuanced, incredibly smart and incredibly necessary because of the stakes that the book confronts that are at play in the world at the moment.

And that's really important. It's very, very, very important.

And it's important that I think everybody reads this book.

It's incredibly accessible as well.

And I said to Jack, who's the director of this podcast before we started recording that there's so many terms like nanotechnology and all this stuff about like biotechnologies and quantum computing at reading through the book.

Suddenly I understood what they meant.

And these had been kind of exclusive terms and technologies.

And I also had never understood the relationship that all of these technologies now have with each other and how like robotics emerging with artificial intelligence is going to cause this whole new range of possibilities that again, have a good side and a potential downside.

It's a wonderful book and it's perfectly timed.

It's perfectly timed, wonderfully written, perfectly timed.

I'm so thankful that I got to read it and I highly recommend that anybody that's curious on this subject matter goes and gets the book.

So thank you, Mustafa.

Really, really appreciate your time.

And hopefully it wasn't too uncomfortable for you.

Thank you. This was awesome.

I loved it.

It was really fun.

And thanks for such an amazing wide-ranging conversation.

Thank you.

If you've been listening to this podcast over the last few months, you'll know that we're sponsored and supported by Airbnb.

But it amazes me how many people don't realize they could actually be sitting on their very own Airbnb.

For me, as someone who works away a lot, it just makes sense to Airbnb my place at home whilst I'm away.

If your job requires you to be away from home for extended periods of time, why leave your home

empty?

You can so easily turn your home into an Airbnb and let it generate income for you whilst you're on the road.

Whether you could use a little extra money to cover some bills or for something a little bit more fun, your home might just be worth more than you think.

And you can find out how much it's worth at Airbnb.co.uk slash host.

That's Airbnb.co.uk slash host.

Thank you.