Are you uncomfortable talking about this?

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

Mustafa Suleiman

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 off 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 question is that

Do you ever get sad about it?

Yeah, it's intense

The show gets bigger which means we can expand the production

Bring in all the guests you want to see

And continue to do in 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

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
With 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
Is 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 hadn'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 eureka moment like a day

When something happened and it caused that epiphany I guess

Was it?

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

At 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

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 at 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, okay, 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 to audio too

What I said 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

Your chat GPTs

Chat GPT

Google Bard

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

So 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 two 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 quite 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

It's my 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?

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

Over the last few years, I would say the default reaction has been to avoid the pessimism and the fear

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

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

I've become used to talking about that and 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

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

I can't see how, I can't see how it can be contained

Chapter three is the containment problem

You give the example of how technologies are often invented for good reasons and for certain use cases

Like the hammer, which is used 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

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 it, like the nuclear bomb, because if they have it, then we don't, then we're at a 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 tumours 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 civilisation

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 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 crossed your arms
No, I always lose 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 mine
It's one of the pinnacle companies that pioneers in this whole space
So you know, you know some stuff
And it's funny because when I watched an interview with Elon Musk and he was asked a question similar to this
I know he speaks in a certain tone of voice
But 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 to stop it
There's one interview that's chilling
And I was filming dragons the other day and I showed the dragons the clip
I was like, look what Elon Musk said when he was asked about what advice he should give to his children
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, that's 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
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

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?

What potential implications?

I think the darkest scenario there is that people will experiment with pathogens

Engineered, you know, synthetic pathogens that might end up accidentally or intentionally being more transmissible

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

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

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
But that is the age we're in
Deal with it
We have to confront that reality
That means that we have to approach this with the precautionary principle
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
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 ten 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 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
Okay, that's powerful
I would like one of those
There might be an adversary out there
Because 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, okay, 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
I 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 the 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...
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 OpenAI 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-3, 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...
60, 70 times smaller than the original model
Which means that it's cheaper to use and cheaper to run
And that's, as we've said earlier
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
Do you believe...
Because we're both going to be alive
Unless 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 has now taken my job
It's very possible
So I'm going to sit you here and you know
When you're...
You'll be about 60, 68 years old
And 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
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...
It's certainly possible
That we invent these AGI's
In such a way that they are always going to be
Proveably...
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
I don't think we should take the gamble
Basically
So...
That's why I think that we should focus
On containment and non-proliferation
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
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
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're never
They weren't ever meaning to do harm
Well, I think I asked you
When I said it, 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 1% difference in DNA
Or something crazy
And yet they throw their poop around
And I'm sat here broadcasting around the world
There's quite a difference in that 1%
You know
And then I think about this new species
Where, as you write in your book in chapter 4
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
AIs 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 are
You know what I mean, what if it replicates in a way where
Because I've heard this conversation around
Like, it depends how we design it
But, 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 could have any influence 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 the meat in my skull can do
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...
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
It's like, it seems likely that if you have one like that
It's going to take advantage of infinite amounts of data
And infinite amounts 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 actually saying no sometimes
It's saying no
And that's a different sort of muscle
That we've never really exercised as a civilisation
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 shifts for me

Because I think, like many people, I thought of artificial intelligence as, you know, like a child I was raising

And I would programme, 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 DeepMind, wasn't it?

That was us, yeah

You guys, so you programmed 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 learnt how to beat the best player in the world of all time, the world champion

He 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

It's a move that seemingly makes absolutely no sense

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 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?

Where we're looking at that move, it's called like move 37 or something

Correct, yeah

Is it move 37?

It is, yeah

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 forecast it
The commentator actually thought it was a mistake
He was a pro and he was like, there's definitely a mistake
The AlphaGo has lost the game
But it was so far ahead of us that I knew something we didn't
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
Like when it discovers a new strategy or it invents a new idea
Or it helps us find like a cure for some disease
That's why we're building it, right?
Because we're reaching the limits of what we as humans can invent and solve
Especially with what we're facing in terms of population growth over the next 30 years
And how climate change is going to affect that and so on
Like 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 and their shareholders are, 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, 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 realisation, exactly what you've just described
Like weighing up two 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 realise
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 alright
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
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
It has to be
You want it to be
I desperately want it to be, yeah
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
Okay
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?

So like, I've got my dial-up modem over here

And then quantum computing over here

Right

What's the difference?

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

1.4 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 gonna have billions and billions of smartphones

In miniscule wearable devices

There'll be cheap fridge magnets that, you know

19/49

Are constantly on everywhere, sensing all the time
Monitoring, processing, analysing, improving, optimising
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
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 gonna be working
Why would we? Yeah, what?
We wouldn't be working in that kind of environment
I mean, look, 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
And 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 gonna 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 gonna power new forms of transportation
It's gonna reduce the cost of drug production and healthcare, right?
So all of those gains, obviously there'll be a huge commercial incentive
To drive the production of those gains
But the cost of producing them is gonna go through the floor
I think that's one key thing that a lot of people don't realise
That there's a reason to be hugely hopeful and optimistic about the future
Everything is gonna 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 gonna be able to upload ourselves to a silicon substrate, right?
A computer that can hold the essence of what it means to be Steven
You in 2200, well, could well still be you by their reasoning
But you'll live on a server somewhere
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
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?
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 gonna be able to emulate the entire brain
So their thesis is that, well, some of them cryogenically store their brain after death
Jesus
So 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 then transfer you over
It doesn't seem plausible to me
When you said at the start of this little topic here that it must be possible to contain it
Said it must be possible
The reason why I struggle with that is because in chapter 7 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 any more
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, well, 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, lifespans 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 rock-riler 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
In a world where there's the ability to generate audio and language and videos
That seem to be real
What can we trust?
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
Because this really brings into focus that our lives are built on trust
Trusting the things we see here and watch
And now 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
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, I'm sure many of us get text messages
I 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, 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 have tracking algorithms that look for 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
In Go it's still...
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
You know, 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 cynical, 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, 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 super humans 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 realise
Yeah that's what I'm saying
I was fine before I started talking about AI
Now then why have I 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
Which is titled
Containment must be possible
I love how you didn't say it is
Containment must be
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
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
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
In their defence 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 just from a personal perspective
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
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 ten years
And so when I say regulation is not enough
What I mean is it needs movements

It needs culture, it needs people who are actually building
And making in modern creative critical ways
Not just giving it up to companies or small groups of people
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
So I have no other than the long term 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 question, right?
The world is happening to us on a decade by decade time scale
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
But 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
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 US
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?
Or why would I want to reduce my AI development
When it's going to make us 15 trillion?
Yeah
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 ten 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?
But have power access to the
Full 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 need 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 the listings with Huell
So you can now get the RTD ready to drink
Tesco Express is 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
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
To 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 end points 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
Super computer in the world today
At least just for this next six months we will
Other people will 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
Couple of years time
That'll be the price of an iPhone
That's the problem
Everyone's gonna 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 fund the massive changes in society
Such as paying for reskilling and education
You put massive tax on it over here
I'm gonna 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
Or Portugal
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 close 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
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
Horse power 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
I don't consider myself to be a pessimist in my life
I consider myself to be an optimist, generally
And 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 many incredible things
And overcome so many things
And I also think I'm really logical
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
Than 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 sieves and stuff
And then we realise the unintended consequences of that

Hopefully not before it's too late
In chapter 8, 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 recognised as the first and the best
They're 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 all want that, right?
There's a lot of people that maybe aren't 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
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 mid term
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
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
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
But 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 pandemic, 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
Now you catch me in a moment where I feel angry
And sad and pessimistic because to me
That's like a straight forwardly obvious conclusion
That 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
Must
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'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?
Well, 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, right?
They have, you know, GPS control
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 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, unlike a car
That has 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
I struggle to believe that you're
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 and near term
And then the problem is, in the short and near term
As is the case with most of these waves

Is we
We wake up in ten years time and go
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 thirty 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 would love to know
Someone as smart as you
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
On the face of it
I care about
I care about science
I care about facts
I care about describing the world
As I see it
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 twenty-first 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 I agree with you
That 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
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
The way that we affect people
The way that we affect 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
Attends to the importance of doing good
In the world
To me that's a 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
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 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
But I feel lucky to have the opportunity
To try and do that
There's this wonderful thing in my favourite
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?
I don't know
What's the range of emotions
That you contend 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
I mean, you know, 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, you know, 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, you know, 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
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
Twelve 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
What I meant by that was
This is a world that we're kind of already in
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 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 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
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 asking
What should I do with my future?
What should I pursue in the light of everything you know

About how our 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
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
Present 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 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
You know, 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 will
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
Are 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 the 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
But 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?
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
You know, opportunities
To quest
You know, 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
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
Everyone gets a check every month
I don't think it's going to quite 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
Like imagine if you did have
Basically free energy
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
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 centre
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
That 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
And
Well in some ways the downside of that
I think failure is not
Achieving a world
Of radical abundance
In my opinion
And more importantly failure is
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
What does that lead to?
Will potentially use that power
To cause harm to others
This is part of the challenge
In this networked globalised 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 destabilise

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
It's a 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
Or what do they call it? Playing to the orchestra
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
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
Biotechnologies and quantum computing
Reading through the book
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
About the 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
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 realise

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
That's Airbnb.co.uk
Slash host