Inside the Mind of Anthropic CEO Dario Amodei | The Circuit | Extended Interview

69 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=X2VHFGYAWPE
https://www.youtube.com/watch?v=x2VHFgyawPE

SUMMARY

The interview discusses the rapid evolution of AI technology and its implications for society, focusing on the experiences and perspectives of Anthropic's CEO, Dario Amodei. He reflects on the pressures of leading in the AI space, the importance of rational decision-making amidst chaos, and the ethical considerations of AI deployment, particularly in military contexts.

- *Amodei emphasizes the need for calm and rational decision-making in high-pressure situations, avoiding paranoia while addressing risks.*
- *He describes the exponential growth of AI and its potential to disrupt traditional software industries, urging companies to adapt or face significant challenges.*
- *The conversation touches on the balance between advancing technology and maintaining ethical standards, particularly in military applications, where Amodei advocates for human oversight.*
- *Amodei shares insights on the cultural influences of growing up in San Francisco and how they shaped his nonconformist worldview.*
- *He discusses the importance of aligning business models with values, particularly in the context of AI's societal impact.*
- *The interview highlights the challenges of maintaining company values as Anthropic scales, emphasizing the need for a strong organizational culture.*
- *Amodei acknowledges the potential for job displacement due to AI advancements but stresses the importance of proactive policy responses and adaptation strategies.*
- *He expresses optimism about the future of AI, believing it can lead to significant societal benefits if managed responsibly.*

How much are you sleeping?
You know, I've never been
someone who slept all that.
Well, let's just say I'm, you know, I'm,
I'm learning the art of, of,
you know, finding ways to relax
and sleep through, through
moments of unusual pressure.
It is all moving so fast.
How does it feel on the inside?
It's this feeling of like the

exponential, like, you know,
suppose you were to accelerate
away from earth on a
spaceship at relativistic speed.
The way special relativity
works is, you know, you go,
you go to sleep and you wake up in two
days have gone by on earth.
And so you have to deal
with two days in one day.
And then you go to sleep.
And then because you've
continued to accelerate,
three days have gone by on
earth, and then the next day
and four days have gone by.
And that's a little
bit what it feels like.
I mean, do you go to bed
constantly paranoid about
what you'll wake up to?
There are enough clear
and present issues that we
have to deal with that I'm,
I'm constantly dealing with
those, while thinking about
how we can prepare.
But, but you know, I, you
know, I, I don't think paranoia
or worrying about what you'll
wake up to is productive.
You know, I've looked at people
in history who've, you know,
who've dealt with these very
high pressure situations
and, you know, you need to
learn to respond rationally and,
and not put dangers out of
proportion to each other.
This yo-yoing between, I'm not worried
and oh my God, we need to panic today.
I, I I, I, I think that's a hallmark

of immature decision making. And
the actual mature decision making
is we can't ignore this.
We can't be complacent.
In fact, it's getting to be
a bigger and bigger risk.
But, you know, we, we have
to respond rationally.
You know, like a, like
a surgeon would deal
with an operation
or, you know, like a military
officer would, you know, deal
with a military operation
or, you know, someone making
decisions that affect a lot
of people has to make
those decisions rationally
and they have to understand the risk,
but they, they, you know, they have
to maintain a basic sense of calm.
So my son yesterday was like,
can I use your Claude Cowork account?
And I was like, absolutely
not. I need my tokens.
We're seeing more and more of them,
even in the consumer space.
We wanted to be more of
an enterprise company.
But, you know, it's even,
even even consumer without us putting
that much effort is starting to go fast.
You are at the center of
the AI universe right now.
What does that feel like?
The interesting thing is, is
that the experience I've had
for my whole career
and certainly the whole
time at Anthropic is
that there's this kind
of smooth exponential

and the experience, the
smooth exponential is,
nothing's happening, nothing's
happening, nothing's happening.
A little things happen and
then zoom, it goes crazy.
That's the experience of the world.
That's the experience of the
scale of the company compared
to the other companies
and compared to the world.
So, you know, I was watching
this graph for a while
and I said, oh yeah, we'll
probably become the AI company
with, you know, the, the most revenue
and the most valuation some
sometime around this time.
And, and indeed, indeed it has happened.
So in one sense, I'm not surprised
'cause there's just a
smooth line on the graph.
But of course in another sense,
when things actually happen,
you just, you see so much
more, you know, detail
and color to it and it, you know,
it definitely is surprising.
And what we're just keeping in mind,
all the things we usually
keep in mind, which are,
which are just, you know,
how do we train good models?
How do we put them in good products?
How do we make sure
that everything's safe?
How do we help people
but also manage the societal
risks around the technology?
It's, it's all the same
questions, just kind
of under a bigger, under a

bigger microscope as it were.
What were you like as a kid
growing up in San Francisco?
I know your, your dad
was a leather craftsman.
Your mom worked in libraries.
How did that shape you?
You know, the whole, you
know, like first, you know,
internet revolution
was happening around me
and I had absolutely no interest in it.
I was just interested in like, doing math
and like scra, you know,
scrawling things, I was interested in,
like understanding the universe.
I was interested in science fiction.
Like that was, that was kind
of the, you know, that was the,
that was the general, that
was the general milieu.
I, I think I just felt a lot
of curiosity about the world.
You grew up in the town where, you know,
that is the center of technology
and right now it's the
center of AI you know?
Is there anything about
this place, this city here
that informed your worldview?
Yeah, I mean, I think
the general, you know,
the general spirit of kind
of, you know, nonconformism
and individualism and
it's okay to be crazy.
I think, I think a good
deal of that probably did,
probably did probably
did rub off against me.
You know, you hear these
stories about, you know, you go

to countries in Europe
or you know, even other parts
of this country where it's,
it's just, you know, it's
just kind of discouraged
or considered weird to like,
think about things in
some different way, right?
Or have some set of, of
some set of crazy ideas.
And, you know, there's a lot
of things I'm actually very
critical about with, with Silicon Valley.
But one thing that I think
is good about it is this,
this encouragement of like, you know,
it doesn't matter if all
the experts are against you,
it doesn't matter, you know,
if you have a coherent
vision and a coherent
world view of the world,
you should go and pursue it.
Maybe it just won't work at all.
But, but if it does, there's
this kind of long tailed ness
to it where, you know, there
are certain places you can,
you know, you can, you can
search certain veins of war
where, you know, you
might, you might find a
huge goldmine there.
I think that spirit is very important.
You, Daniela, your sister
and her husband, Holden
Karnofsky lived in a group house
together back in 2016.
What were you debating back then?
That was, I think the
time when, you know,
Open Philanthropy project

was, was, you know,
first being startup, which
Holden was the lead of.
And I was at that time, you know, like a,
a biological scientist.
So, you know, I was helping them with some
of the stuff they were doing around kind
of developing world health
or biological research.
So, you know, I I kind
of advised on that stuff
and, you know, what were the
areas that were promising?
What were the areas
that were less promising
Your decision to leave Open
AI has become Silicon Valley lore.
What really happened?
Like, beyond the narrative,
what were the issues?
What did you disagree on?
Look, I'll, I'm gonna say it,
I'm gonna say it very simply.
You know, there are many
difficult issues that you know,
you face when you're
building powerful technology
that Anthropic faces every day
where we don't know whether
we're making the right
decision or the wrong decision.
So, you know, there are
many valid disagreements
to be had on safety.
We certainly had some of
those disagreements with them,
but, you know, people that, that, that,
that alone is not sufficient to leave.
People here have had
disagreements with me,
people here have
disagreements with each other.

But when you feel that
you can't trust someone,
when you feel that their values are not
what they say they are, when you feel
that they're not honest, when you feel
that they're not in it for
the reasons that they say,
when you see disturbing patterns
of behavior dishonesty,
that makes it very hard to,
you know, to continue
to work with a company,
to continue to trust the company.
And look, at the end of the day, why argue
with someone when you don't have the same
vision and you don't trust them.
Like the way the the way
to resolve it is you go off
and do your thing.
They go off and do their thing.
And I am completely at peace with the idea
that we're doing things our way
and they're doing things their way.
We'll see who wins in the market
and we'll see who wins in
the court of public opinion.
I think those things speak
louder than any drama about why,
who left, what, you know, we're,
we're providing an example
of how to deploy this
technology, you know, in
what we think is a responsible way.
If they disagree, they
should make that argument.
And you know that, I think
that's really all there
is to say about it.
There was a moment at
India's AI summit where you
and Sam Altman appeared to refuse
to hold hands on stage.

What happened there?
What happened is that the summit was
extremely disorganized.
We all came up at the last minute
and they like changed the order
in which we were standing,
and then like, they took a picture of us
and then they ordered us
all to like, hold hands.
You know, if you've ever
been to one of these summit,
I am not saying anything bad
about India in particular,
but like all of these kind
of international type summits
that have like heads of state
are like super disorganized.
Okay? But everyone
else held hands. Come on.
I, I look, I don't know, I
don't know what to tell you.
Okay. There was like, you know, Narendra
Modi up there suddenly
telling everyone to, like
suddenly telling everyone to hold hands.
All right, all right,
well, okay, look, Sam
and Elon are suing each other.
You don't like Sam. I,
it seems if the people
building the most important
technology in the world
can't hold hands on stage,
how can we trust you'll
cooperate on existential risk?
So here's, here's what I will tell you.
There is a wide variance in the quality
and the trustworthiness of the people
building this technology.
I think this meme that,
you know, different,
that no one trusts each other,

I don't think it's right.
You know, I've known Demis Hassabis
who builds the Gemini models. They're
a competitor to Claude models.
I've known him for 15 years.
We've worked together on like,
you know, a number of issues.
We buy compute from Google,
we swap safety ideas all the time.
So, you know, my my view
of this is that one,
there are some players who are
more trustworthy than others.
And you know, I think there are players
outside Anthropic who,
you know, who, who I trust,
who I see as trustworthy.
What I think needs to happen is
that the trustworthy actors
need to need to get together
and, and put the untrustworthy
actors in a position
where they kind of have to
adopt the same standards.
With a lot of experience, I've learned
that there are some folks who
don't do the right thing on
their own, but if there's
a majority of the industry
that's doing the right
thing, then I think the rest
of the industry is, is kind
of, they're left in a position
where there's not much
they can do that, that,
that then come along.
There's like the positive version of it
where you inspire other
people, that's like Demis
and me inspiring each other.
You know, he does Alpha Fold.
We're trying to do something

in bio as well, right?
We do interpretability research,
they start an interpretability research.
It's not even competition,
it's just, it's just, you know,
each company does something cool
and the other company's like, that's cool.
We'd like to, you know, do that too
and see if there's something
new within that we can do.
So that's the kind of,
you know, the carrot side
of the race to the top.
Then there's the stick side
or the implicit stick
where you're like, okay,
these guys are doing the right thing.
Those guys will look bad if
they don't do the right thing.
And often we see behaviors where they kind
of grudgingly do the
right thing while trying
to pretend they're doing
something different.
And there's something
bad or sinister about us
that is to be expected.
But I think that's the way
we get the industry together,
and that's the way we get
the industry to cooperate.
Now, early on, others focused on fun,
splashy consumer apps.
You made a bet on coding and enterprise
and Claude Code is a hit.
Claude Cowork is a hit.
Why did you make that bet?
Was it a values decision
or a business decision?
When we started Anthropic,
the thing that the,
the base thing that mattered, the thing

that always matters is we wanna,
we want to do this, right?
But then you have to ask
yourself, okay, in order
to fund the very expensive,
you know, creation
of these models, it, it
needs to be a company,
it needs to have a business model.
Does the business model get
in the way of the values?
There's always this question,
but I think one of the things
I learned is, you know,
just from being at other companies
and watching other companies is, look,
if you pick a business model
that fundamentally conflicts
with your values, you're
gonna have a hard time, right?
Either you betray your own
values or you become irrelevant.
You know, you kind of end
up in a catch 22 situation.
And there are ways out,
there are ways to dodge,
but it's just, it's just a hard situation.
It's far better to pick a business model
that is compatible with your values.
And so when we thought about
it, we said, look, you know,
we've seen the world of social
media, the consumer world,
it, it really seems to, you
know, encourage engagement,
even addiction, you
know, the slop we've seen
with AI video models, it's
like, what's going on?
It wants to maximize the
number of minutes that you're,
you're paying attention to,

because that's the advertising
revenue driven incentive.
Whereas if we look at enterprise,
look, I mean, you know,
we want to make these
models useful to people.
If I think of all the positive things you
can do with AI, right?
I warn a lot about the negative things,
but ultimately we think
the positive things will
outweigh the negative things.
Many of those are basically fall under the
banner of enterprise.
You know, we want to use AI
to, you know, cure diseases
that we couldn't cure before, right?
Well, that's working with
biotech, it's working with pharma,
it's working with
academic research groups.
All of those are enterprises, right?
We want to use AI to like, you know,
to make energy cheaper and more efficient.
That's, that's all enterprise.
You know, we want to use
AI to help with education.
Most of that is enterprise.
You know, we want to use AI
to, you know, to address,
you know, health and developing
world. Well there are nonprofits.
But those are basically enterprises.
We want to increase economic growth.
That that is basically enterprise as well.
And then I think there's
another factor, which is
that enterprises care a lot about trust in
long-term relationship, right?
Consumer can have this, you know,
almost this gimmicky aspect to it, right?
Where with enterprise, it's like,

what matters is you build a
relationship where, you know,
you work with, you work with
a company for many years,
you know, you deliver on
what you say they deliver on
what they say, and they
basically trust you.
And so it's very synergistic
with our goal of, you know,
deploying these models in
a positive and safe way.
And so I think it serves us
well to have this business model
that largely aligns with our values.
Not that there aren't conflicts sometimes.
Not that there aren't hard
choices we have to make,
but I think the number of such choices,
it's much lower than
it would be otherwise.
A developer can switch
from Claude to ChatGPT
or Gemini in an afternoon.
Is it really possible
to have a long-term lead in this industry?
And, you know, how long would
it take a serious competitor
to replicate what you've built?
Model quality is the
most important thing.
Like we're, we're very far ahead
right now on model quality.
There is some amount of inertia,
but I've never relied on that, right?
I've never relied on
like the, you know, the,
Anthropic has never
relied on like, oh,
this is sticky and people won't switch.
I think you wanna have a better model.
You want to have a better product.

And you know, we, we see the growth rates
haven't inflected at all.
If anything, they've gone up
at least, at least at the time
of taping this interview.
So, you know, I think I, I I tend to think
that is the most important thing.
Soon after Claude Cowork was released,
$285 billion in market
value vanished overnight,
traders called it the SaaSpocalypse.
If AI continues improving
at this pace, how much
of traditional software
gets replaced and how fast?
Yeah. So, you know, this is, this is one
of these questions, it's kind of very hard
to predict in advance, right?
If you could predict it
perfectly in advance,
then people would,
and they'd make a huge
amount of money on the market
and they'd always be right.
So, you know, no one,
no one knows exactly
what's going to happen.
But I would note a few things, right?
All of these traditional
software companies have
a number of moats.
I think what's gonna happen is some
of these moats are gonna go away,
but others are going
to stay around, right?
The ability to quickly write software, I,
I definitely think
that's going away, right?
If, if your moat is, we
wrote this complex software
that no one else can
write, like good luck,

you're not gonna be able to defend that.
But I think folks have
customer relationships.
Folks have know how of how, you know, of
how the field works.
Folks have unique domain knowledge.
So I think my advice to all of
these folks is, obviously,
you know, don't be
complacent, don't ignore it.
Make a list of all your moats
and be very aware that some
of them are going to go away,
while others are going to become
relatively more important.
Because there are limiting factors
and there may also, there
may also be new moats.
And I think those that deftly
respond, that, you know,
lean into the list of moats
that are still present as well
as the new ones will do well.
I think those that are complacent,
that kind of, you know,
just delude themselves at
what worked in the past will,
will continue to work.
They're, they're not gonna have a good
time.
So that is, that is
the advice I would give.
And you know, I I, I, I
think at the end of the day,
I would guess, I mean, it
depends what you call SaaS
and what you don't call SaaS,
but like, I would guess
that the software industry
gets larger, not smaller.
Although there, there
will be some big losers.

Explain that.
I just think the pie
is getting bigger, right?
Like, I think, I think with ai,
like the pie is getting bigger,
the existing incumbents may
be smaller in relative terms.
Some of them may, may
go down in value, some
of them may even may even go out
of business if they don't
adapt in the right way.
But I, I, I, you know, I think you,
I think you see this often when growth is
really fast, right?
If the, you know, if the,
if, if, if what's possible
with AI grows by 10x, it's very easy
for an existing incumbent
industry to go up by 1.5x, right?
Just, just, you know, not as much
as the whole big pie is growing.
So I think that may happen that that's not
to say we won't have such some big losers.
I think those who don't adapt,
who put their heads in
the sand, who don't kind
of see what's coming, who
don't identify the moats they
have, they're gonna
have a really hard time.
Your biggest backers are
companies like Amazon and Google
and Microsoft and Nvidia.
These are companies that
all have their own agendas.
They are partners in rivals.
You have huge commercial
milestones tied to funding,
who's really calling the shots.
There have been a number of cases
where we've really spoken our

minds about what we think,
you know, I've been very
outspoken about the need
for export controls on
ships to China, right?
I, i, I say this
because I think it would be
really bad for America, for,
you know, the state of democracy
in the world for, you know,
China to be ahead in AI capabilities.
And, you know, it's, it's like some
of the chip makers obviously
don't agree with that view,
but it hasn't stopped me from saying it.
I'm saying it again now, even
after we've signed more
partnerships, what they know is
that we always work with them.
We've been good partners, you
know, we can work together.
I'm sure they wish we
didn't say these things,
but these things are
what I believe.
What are you gonna do? You know,
they're, they're at the end
of the day, they want the, you know, they,
they benefit from these
deals as much as we do.
You know, look, we're all adults here.
We can work together on one thing while
disagreeing about another thing.
Bloomberg's reported
that you're at valuations
that are higher than OpenAI.
We're talking nearly a trillion dollars
for a five-year-old startup.
How do you make sense of that number
and why do you need that
much money if you, you know,
you're more disciplined on compute,

you have a faster path to profit,
The compute is ramping
up very quickly, right?
So it, it can both be the
case that the fundamentals
of the business look good,
but in, you know, in a year
you'll have three times
as much compute as, you know,
three times or four times,
or, I'm not gonna give exact numbers,
but like these compute
ramps are very fast.
And we have every expectation
that the revenue, you know,
ramp will meet and exceed those.
But raising money is, is kind
of the buffer against
this cone of uncertainty.
So it's a totally rational thing to do.
It's, it's a very small
dilution to the business,
and it logically is not
at all the same thing.
In fact, it's compatible with
the opposite as you know,
that there's anything wrong
with the fundamentals of the business.
There have been
reports of server strain,
reliability issues,
people complaining about
running out of tokens.
You've said other companies are
yolo-ing on infrastructure.
Do you actually have what you need
or are you playing catch up?
So one of these things
about compute is there's a
marketing compute, right?
So, you know, my view is
that over a period of time,

even longer than a couple
months, like, you know,
we can get large amounts of compute.
One, one thing that's worth
saying here is, you know,
I don't think we bought too little compute
by any reasonable standard.
So, you know, we were planning
for a 10X a year growth in compute,
10X a year is what we expect.
That isn't what we've seen
over the first quarter of 2026.
We saw a greater than three x
growth in revenue quarterly,
just in not annualized three
x in the quarter, which
of course, three to the
fourth power is 80x
over the course of the year.
We didn't plan for 80x annualized growth.
It would not have been rational to plan
for 80x annualized growth,
because that means if you only
get 10x, you know that you,
you have eight times less.
So we're, we're, we're in a
locally extreme, you know,
explosion of compute
that's not gonna continue.
If that continued, you know,
you just get to revenue
by the end of the year,
you get to revenue numbers
that no company on earth, I
don't think that's gonna happen.
It just, it just can't.
But you can have these short
periods where it's like,
oh my God, like, you know,
this is faster growth than we ever,
ever possibly anticipated.
But I don't know, you saw the

compute deals with Google,
you saw the compute deals with Amazon.
You know, there are
more that we kind of can
and will do, like, you
know, the market's liquid.
Like if, if, if, you know, if you're able
to use compute really well
and there's the demand,
you'll get your compute.
It might might just take a month or two.
Does it feel good to
surpass your arch rival?
Look, I, we have a lot of
difficult challenges in front
of us, and there's this
race to the top idea
that we're trying to pull
other companies along with us.
And I think we've seen that.
We have pulled them along with us.
Sometimes they don't admit
that that's what they're doing.
Sometimes they copy us
while they're attacking us.
But, but this pull is very valuable.
And so I think the value of
being the preeminent company,
both commercially and in terms
of models, you know, it's,
it's not about beating rivals for the sake
of beating rivals.
It's, it's about having the ability
to pull the ecosystem along with us.
And we hope that we can do
more of that in the future.
But winning has to feel
just a little bit good.
I mean, look, we're always
trying to succeed, right?
Like, we're always trying
to, you know, we're not,

we're not trying to fail here, right?
Like, I'm not someone who
believes we should shut
this technology down.
We shouldn't build it like,
you know, we, we, we, we,
you know, we, we exist within
a free enterprise system and,
and, you know, there's, there's nothing,
there's nothing wrong with this.
We just have to mitigate the
risks of the models, right?
And, and so it's always been
the balance between the two.
Now for most of Anthropic's
history, you were the underdog.
I imagine it's easier
to take the moral high
ground when you have nothing
to lose at this scale.
How hard is it to stay
true to your values?
What I would say is that,
you know, I've put a lot
of time into thinking about
how, how that's the case.
You know, as, as, as companies scale.
You know, I've been
paranoid at every scale,
at every scale of the company.
There's some new challenge,
there's some new way the company can lose.
Either its, its kind of will
to win just commercially
or kind of the core of its values.
I'm, I'm worried about both
because I see them as synergistic.
I actually see the fact
that we've been able
to make such good models as
the thing that, that allows us
to assert our values in a way that works

as the company grows, as it gets bigger.
There are lots of pitfalls here.
There are lots of ways to go wrong.
Not because me or the co-founders
or the company's leaders
values change, but
because the composition of
the company changes very fast.
So I spend probably half
of my time just talking
to the company about
the culture of anthropic
and how the culture works, right?
When you're growing this
fast, you're hiring a bunch
of people from big tech companies.
If you don't tell them
how Anthropic operates,
they'll simply recapitulate
the only thing they know,
which is how to operate at the
companies that they came from.
And so this is a constant
struggle and a constant challenge.
And, you know, it's like, you know, me
and Daniela's, maybe
number one top priority is,
is figuring out how to preserve this
because we recognize
that this is the core of
who we are in the long run.
Your product velocity is insane.
You're shipping so much so
fast. How are you doing it?
I would say two things.
The first is, you know,
we have a unified company.
We have a unified culture.
You know, I think we've gotten, you know,
grown larger while still
being incredibly efficient.
Everyone's still being on the same page,

like just the cultural
and organizational unity.
I would say that's the biggest factor.
And I would say the second
biggest factor is Claude itself.
That we're now using
Claude to help, you know,
develop our models and, you
know, make them more efficient
and quickly develop products.
There's all kinds of new
practices you have to develop.
You know, we're, we're still new, new,
we're still new at it,
but you know, it's producing,
it's producing a lot
of acceleration and increasingly producing
reliable acceleration.
And so those are the two
factors I would point to.
Will you tell me the most
wild thing you've seen AI do?
I think some of the
wildest stuff I've seen is
around biology and medicine.
I've seen a number of cases,
including Daniela actually,
where Claude diagnosed a
medical problem that, you know,
a bunch of fancy doctors had missed.
And on the biology side,
like the models are starting
to get surprisingly good at
like, you know, you know,
tasks like drug design or, you
know, computational chemistry
or things like, and I'm just
like, wow, you know, as someone
who used to be a biologist, I look at it
and I'm like, wow, that's hard.
Like, you need a lot
of training to do that.

And like Claude is getting good at it.
And that's one area where
I think we're gonna get a
hell of a lot of benefit.
Like that's the positive for AI.
We're gonna get these
huge, enormous benefits.
Life is gonna get better.
The quality of human
experience is gonna get better.
A century of scientific progress,
A a century of scientific progress
and a century of progress.
And what it's like to be human.
Like a go back to 1900, think
of all the problems we had in 1900s.
All the reasons people died prematurely,
all the problems they had to suffer,
all the material deprivation
that we don't have to deal with today.
Then think of another
hundred years of that.
I really believe this
century of scientific
and medical progress, if
we can get through this,
and I, I think we will.
I'm increasingly optimistic.
We're gonna have a
much, much better world.
I know how much you love writing.
You're known for your essays. Do you
use Claude to help write?
I do. I have not gotten to the point
where I actually allow text
directly written by Claude in,
in, because I, i, i, I just
have such a specific style
that I'm, I'm a little picky about it,
but I basically use
Claude to like, you know,
to help me brainstorm, to help

me think through the themes
to help me kind of, oh, you know,
what are some references
I could use for this?
So it it, it kind of
plays a supportive role.
I don't know how far we
are from Claude being
able to write better than me.
We're not quite there yet.
But, but you know, I think, I
think certainly it's coming.
I love writing too, and I,
you know, I feel like writing
it helps you struggle through ideas.
There is a lot of critical
thinking involved in that.
Do we lose that if we let Claude do it for
us?
I, I, I'm a little worried about that.
And in fact, that's half
the reason I write myself.
It certainly is for external audiences.
Many people read what I write,
but it, it, it is just as much
to clarify my own thinking so
that I kind of know what to do next
and to create a common reference
point across me and others.
I think we're still grappling
with the question of
how exactly do we use
AI in a way that kind
of preserves those benefits.
I think the thing I'm doing now
does that where I use Claude
for research and I use
Claude for kind of, you know,
how do I help organize my own thoughts.
I think if we just used it end to end,
like write an essay about the
risks of ai, first of all,

it wouldn't write the things that I think,
but also I would, I would
exactly lose that benefit.
There's some way, as the models
get better, I think probably
to, to use them directly much
more directly in the writing
and yet still preserve those benefits.
But I think it's gonna be a subtle thing.
It won't be all one thing.
It won't, we'll have to kind
of figure it out over time.
I think we could have this
very unusual combination
of very fast GDP growth
and high unemployment,
or at least underemployment
or, you know, low wage job.
Lot of low wage jobs, high inequality.
You've been really
direct about job loss.
AI could eliminate half
of all entry level white collar jobs
in the next one to five years.
That was a year ago. AI
has moved incredibly fast.
Is it still 50% or is it higher?
I've always said,
and you know, if you go back
to those original clips,
they always get like, you
know, cut out of context
and like the three seconds.
But like, you know, the, the
real statement was always,
I don't know what's gonna happen,
but this is an order of magnitude for
how crazy things could be.
Also, I always talk about
all the things we can do in
response to this, right?
I've talked about token tax

and working with enterprises
to adjust people,
and I'm a little skeptical
of retraining programs,
but like, we should throw them in the mix
macroeconomic policy,
even from the beginning.
I always talked about solutions,
but you know, somehow there's
this tendency in the human
psychology to clip the three seconds
of like, doom is coming.
So my message is just
definitely not doom is coming.
My message is like, this
is something, you know,
that we should see coming,
that we're worried about
and that we need to actually
respond to positively.
You know, I don't know exactly,
but I'm, I'm still pretty concerned.
I'm still the same order of concern.
You know, we are seeing right now
that AI is making people more productive.
But that's the usual hump.
If you go back, you know, to
the kind of industrial,
you know, I wrote about this
in Adolescence of Technology.
You automate 90% of the job,
great, people are 10 times more
productive in the other 10%.
'cause they're 10 times more leveraged.
But eventually it gets
close to a hundred percent.
Now the sequel to that
is, well then you have
to find something else for them to do.
I, I don't know about the long run,
I'm truly uncertain about that.
But I do think there

are types of adaptation.
Like one thing I'll
talk about is, you know,
software engineers within Anthropic.
We're, we're going through this
transition right now where,
you know, right now AI makes
the software engineers more
productive, even though AI writes all the
code or almost all the code.
But still it makes people more productive.
But we're already starting
to see the beginning of like,
you know, there may be
some people that it's,
it's not making more productive
that it's better for the AI
to just, to just do the thing.
So that's one side of it.
The other side of it though is
what do we need more demand for?
You know, there's something
we call a forward deployed
engineer or in like applied
AI solutions architect
where their job is a mix
of technical work and
talking to customers.
There's a lot of demand for that
because there's a lot of customers
and we're growing very quickly.
Now, does every person
who is in the pure
software engineering quite
work for the this other?
No, you know, it's not perfect.
It's not one-to-one.
That gives you a flavor
of, there's gonna be a hell
of a lot of disruption,
but things will also
adjust. Which wins out?

I don't know. But the
reason it's important
to warn about it is that
that's how we can respond.
That's how we can make policy, right?
Both within Anthropic
and macroeconomically for the whole world.
We wanna put out carefully
considered thoughts.
We don't wanna say things
that people don't
believe we'll actually do.
We don't wanna say things
that are half baked.
We want to think carefully about
what should actually be done about
these, these problems.
You put out this chart
showing potential job disruption
like sales, finance, you
know, which jobs go away,
who gets replaced and
what new jobs are created.
So no one knows for sure,
because you know, the
economy's unpredictable, right?
It's the same as the stock market, right?
They're these kind of
decentralized processes that you,
you don't really know ahead
of time what are the pieces
of the job that people are
still gonna be able to do.
But what I would say broadly
is that, you know, anywhere
that you have, you know, these kind
of entry level white collar,
you know, whether it's banking,
whether it's finance, whether
it's, you know, there's,
there's, you know, there's,
there's gonna be a lot

of potential for AI to first
make people more productive.
But you know, then, then, then
there's gonna be, you know,
then, then there's gonna be
a wholesale AI can do the job
and then we're gonna have to
think about, well, you know,
what is it that people can do?
And I think we need to plan
about that ahead of time.
We're already doing it. When we talk
to enterprise customers, we
see choices that they face.
They face the choice of, you
know, should I save costs?
Which often means hiring less people,
basically do the same
thing with less resources,
or should we do more things
with the same amount of resources?
And we always, when we can
try to push them to doing more
with the same amount of resources,
because basically that means
like, hire the same number
of people or maybe even more people,
but just do, do kind of,
kind of do new things.
Pushing them towards the positive sum.
The thing that that we have going
for us here is the pie
is gonna expand a lot.
And so, because the pie
is gonna expand a lot,
there are probably going to
be places where people can go.
It's just a matter of
finding them fast enough.
It's the size of the disruption.
It's, it's gonna be big.
And that's what I'm warning people about.

But we kind of, we have to
solve that matching problem.
And
So play this out for me a little bit.
You know, you wake up in five years,
what does this country look like?
What are those people doing?
Yeah, so 'cause if there's
that much unemployment, is
that not how revolutions start?
Yeah, no, this is the
outcome we wanna prevent.
This is absolutely the
outcome we want to prevent.
You know, I think, I think there's,
I think there's a few places,
none of them are guaranteed.
We're not sure, but there's
the physical world, right?
Like things that are in
the physical world, yes,
there's a robotics revolution as well,
but it's a lot slower than
what's happening in AI.
People always talk about
building data centers,
but like when processing information
of any type becomes a lot easier,
maybe the restriction is gonna be
things in the physical world.
And so we need a lot
of more people to make,
build, manufacture things
in the physical world.
Anything that's human centered, I think
that's gonna be a big deal, right?
I hear all these stories about
AI found something that my,
like my doctor couldn't
find and I feel, but like,
but there's a, people really
wanna talk to other humans,

particularly over kind of
important things, right?
Maybe AI can do better customer service,
but nevertheless people,
or at least some people
wanna talk to humans.
So these kind of human
relationship driven jobs, like,
I think those are gonna
be important, right?
And I think there'll be some
effort by the humans to kind
of direct the ais right?
At, at some level it has to be
in line with someone's values
and someone's intentions and,
and so I, I think there's
gonna be some role there,
although I don't know how thin
versus how thick it will be.
And I think it's very hard to say.
There has been a lot of pushback,
and I know you've said
you're trying to warn people,
but that, you know, you're, you know,
Jensen Huang said you're
conflating tasks with jobs.
Other folks have said
this, you know, it's sort
of doom marketing that benefits Anthropic.
So, so I wanna be really clear
and push back hard against
this, the whole picture
of there are risks to job
loss and here are some ideas.
I mean, we haven't fully
fleshed out the ideas
because I want to get them right,
but Anthropic has come
up with lots of ideas.
We've had economic grants,
we have the economic index.

I talk about the, the possible ways
to address these risks from tax
and macroeconomic policy to
what the new jobs are in the
adolescence of technology.
I lay out, you know, I have
like five pages where I lay out
the difference between tasks and jobs.
Why this time is different
than other times.
A list of six different
things we can do from private
philanthropy to government action.
I talk about the problems,
I talk the solutions,
but social media, which
I detest, which I detest
as a category, people have
these three second clips from,
you know, from a year ago.
They don't actually read the essays
or they prey on the idea
that social media, I've,
I've written much more carefully
about these things where,
where I talk about the risks, the idea
that this is cheap marketing
is itself cheap marketing.
This is, this is laziness,
this is failure to engage
with serious intellectual work.
And, and I think that
is part of the problem.
Again, I think it's, it's part
of the disease of Silicon Valley.
It's been caught up in
this social media world of,
of, of three seconds.
And so people only respond to it
or they think they only
have to respond to it again.
I think it's very dangerous

and we failed to have
a mature conversation.
Instead, people just lazily
see this like three second clip
and, and they're like, oh,
this is what Dario was saying.
It's, it's so stupid. It's so unserious.
And whenever someone
says something like that,
I take them less seriously.
One of the leading AI
companies in the world is deeply
embedded in many different aspects
of US national security
across military operations,
The standoff between Anthropic
and the Pentagon over AI military
safeguards is ramping up.
You've had a longstanding
anti-war stance starting all the
way back to your days at Caltech,
and yet you were one of
the first AI companies
to sign a contract with
the Department of Defense
to operate on classified networks
that the US uses to
fight wars. Explain that.
Yeah, so, you know, what
I would say is, is look,
I mean the world, the world changes.
Like, you know, my my view
of this technology, you know,
when I see Russia invading
Ukraine, when I see the risk
of China invading Taiwan, it
worries me that we have a kind
of resurgent authoritarian block
that they're very aggressive
and that we need to defend ourselves.
That is something that I,
you know, have believed

for a while now continue to believe.
And, and that's why across
both administrations, you know,
you know, across, you
know, I may not agree
with every policy of I,
of either administration,
but you know, that's why we've generally
been supportive of this.
We don't want a world where China
and Russia can build, you know,
can analyze all the intelligence
with AI, can, you know,
can, can use AI for, you know,
for attacking Taiwan and Ukraine.
And, and we can't defend them.
So that's why we worked with them.
We certainly don't do
it for the, the money.
It's a huge pain.
You know, even, even even
putting aside the, the lawfare,
it's just a huge pain to get
up on government networks
for not that much money.
So we, we did it because we
cared about it, but similarly
because we're doing it
because we cared about it, there need
to be limitations on the
use of the technology
and the formulation
that I used in Adolescence of Technology.
We should use this technology
in every way except the,
the ways that undermine
our own values, right?
And our red lines of mass surveillance
and fully autonomous weapons.
Those are things that I
believe undermine our values.
It's not worth democracies winning if

democracies do those things.
And, and so that's the,
that's the balance that I,
that I see, and that's
the stand that we took
and it, it explains both why
we were the first to work
with Department of War and
why there were some things we
wouldn't do when,
when others were willing
to do those things.
I, I think you need to pick a
stand and stand your ground.
This idea of, you know,
companies that seesaw from,
we won't do anything with the government
to suddenly we're doing absolutely
everything with the government.
I don't, I don't get
it. You should pick your
principles and stick with them.
You've been working
with Palantir since 2024.
That's right.
You know, their technology is used
by ICE, police departments, in Gaza.
Is Claude being used for
surveillance in other ways?
We don't work with ICE
either through, either
through Palantir or anyone else.
We don't work with CBP, I
don't believe we work in Gaza.
You know, our, our, we're
we're very careful about,
you know, scoping our engagements
to things that we believe in.
So, you know, you drew
your, your red lines,
the president banned you
from the federal government.

The Pentagon labeled
you a supply chain risk.
OpenAI jumped in
and signed the contract that you wouldn't,
what does winning this
fight actually look like?
You know, I don't think
there's any winning.
This fight for a private
company like this isn't a fight.
Anthropic is, is trying to win
or thinks about winning or losing.
This is more a, I won't
even call it a fight.
This is more a debate about
what the proper use of AI
by the government is.
And AI is an emerging new technology.
We don't understand the ways in which it's
reliable or unreliable.
We don't understand the ways
in which it promotes our values
or undermines our values.
And so one of the things that
I thought was important was
to establish a precedent
on some of the, some
of the use cases we think are
good, which frankly is most
of them, and some of the use cases
that we're concerned about.
And as I've said, we've
already seen, you know,
you can only do so much
with a contract, right?
As we've seen someone
else can sign a contract
that doesn't respect
your, your same red lines.
But what it has done is raised
awareness for the issue,
and then we have serious

bipartisan efforts in Congress
attempting to ban some of the things
that we're concerned about and
attempting to set guardrails.
Again, I don't want to
talk about this as a fight,
but that's kind of winning the effort
to get our country to
think more carefully about
what is appropriate
use of this technology.
That's. Anthropic is run
by an ideological lunatic who
shouldn't have a, sole, that's,
But.
That's not my.
Question. My question is
AI decision making over what we do,
Do you mind being called
an ideological lunatic
or a bunch of left wing nut jobs?
You know, I've been called worse things
than that all the time.
You know, people, people can call me
or, you know, people can call me
or anthropic whatever they want.
The two things that matter
are, we're successful
as a company and, you know,
we stand up for our values.
Like, I actually, in some
ways my life is really easy
because when those are your, you know,
those are the two things
you're trying to do.
It's, it's, it's really simple, right?
Like it, you know, you just,
you always know where you stand.
A US official has said
with the help of LLMs,
the US military has gone from being able
to hit a thousand targets a

day to 5,000 targets a day.
That means Claude can help
kill more people more quickly.
Are you comfortable with that?
I think there's two things here, right?
There is, there is the
ability of the United States,
you know, to be more effective militarily.
I, i I am supportive of that ability.
I think having that ability be
stronger doesn't cause wars.
It deters wars.
Like, you know, I basically,
you you're asking like,
you know, do you believe
in this country, right?
Do you want this country
to be a more powerful actor
rather than a less powerful
actor on the world stage?
I do, I'm a patriot.
There's a separate question,
which is, you know,
are there particular policies
that the US government is
engaged in that I might support
or not support?
Obviously I support some of them
and I don't support others of them.
It's not up to me. If we
provide a technology, you know,
the DOW made this point and
we actually agree with them.
If we provide a technology,
it's not up to us to say,
you can do this military operation
and you can't do that military operation.
Now, I might privately believe
that this military operation makes sense
and that military operation is a bad idea,
but we're not gonna deny the technology.
You, you have to leave policy in the hands

of the military decision makers.
What you can do is to assert
some high level boundaries
that, you know, for, for
us, prevent the use cases
that seem inconsistent
with, with our values,
with our country's values,
and promote the use cases that
we think, you know, we think,
we think encourage our values.
So that's how we think about it.
Bloomberg has reported
that Claude is being used
by the US military in the war in Iran
to do AI assisted
targeting via platform made
by Palantir, Maven Smart
System in February.
A US missile reportedly hit a
girl school in Iran killing more
than 150 people.
Most of them children.
Did Claude play a role in that strike?
We look, we don't have
access to, you know, we,
we don't know exactly how, you
know, these models were used.
You know, obviously like, you
know, these things that, that,
you know, mistakes that
happen in warfare are
really, really terrible.
Like, this is a really
terrible thing to happen.
If that doesn't make clear
why we have to, you know,
stand up for use cases
that, you know, we don't,
we don't support, like, you
know, we, we, we were willing
to risk the future of our

company to like limit how,
you know, these models are used
and you know, what, what you're
talking about is a use case
that doesn't even violate our red lines.
We're worried that there will
be a hundred times as much,
you know, with, with use cases that, that,
that do violate our red lines.
Now I, you know, I I I,
you know, again, again,
I would say I think overall
the use of these, the, the use
of these models is appropriate.
I think it's good on net, you know,
but military decision makers
make terrible mistakes even
when, even, even at the best of times.
And I, I don't know if
we're in the best of times.
Like there's several
things we can talk about.
We can talk about making
red lines that, you know,
prevent uses of the models
that are more likely to lead
to those problems, right?
If we had allowed, you
know, fully at high,
if we had just given, in which
almost every other company
now has to fully
autonomous weapons, right?
This is like a human, what
what we've seen here is Claude assists,
but a human makes the final call.
So a human made that
final call, not Claude.
Imagine if you had a world
in which, not Claude,
because we haven't allowed it,
but someone else's AI model.

The AI model just makes the decision
and the human never sees it.
That's what we were standing up for.
That's what we were fighting against.
I would, I would also say, you know,
there's a separate thing here.
Again, I don't think procurement
is the right way to do it,
but like, you know, we, we,
we need to make sure that,
you know, it's, it's a matter of interest
to the American people,
not to me as a supplier
of the technology, but
to the American people
that are military decision
makers don't make these mistakes
that they operate
reliably, that, you know,
they choose wisely what to do.
Again. You know, that's, that's
of concern to me as a, as a,
you know, as a citizen, as a
supplier of the technology.
Like, you know,
the government uses
Microsoft Excel a lot.
You know, if I said micro, you
can use Excel for, you know,
this military operation,
but not the, you can't, you
can't realistically do that.
But hopefully that gives you a
sense of how we think about it.
This school had a website.
You could have found
it in a Google search.
Like shouldn't Claude have
spotted that, shouldn't AI
or whatever technology they
used have spotted that.
And does it speak to a scarier

issue about using technology
as a shortcut in war.
Look, I look what
I'm, what I'm, you know
what I'm gonna say is, you know, and,
and I, you know, I don't, I
don't know, this relies on,
you know, maybe classified
knowledge that I don't have.
But you know, the principle
that, that we have established,
and I think the principle
that was obeyed here is
a human makes the human
makes the final decision.
I don't know what role
Claude or any other AI had,
but like, if, if this
isn't an illustration why
that principle is so important,
I don't know what's, is
Is AI warfare more likely
to stop World War III, a war
between the US and China?
Or is it more likely to make it happen?
I would say on balance it
is more likely to stop it.
But if we have no limits on how it's used,
then I think, you know, it could
be more likely to cause it.
You know, you've seen
Doctor Strangelove, right?
The premise of it was like
you have a doomsday device
that automatically fires
nuclear weapons when it thinks
nuclear weapons are being fired at it.
What could go wrong? Right.
Again, i I I get to this lethal, you know,
fully autonomous weapons thing.
I think the way conflicts
happen is that, you know, the,

the two sides jump at each other.
They misunderstand each other.
And when we don't have proper
oversight of this technology,
I think those kinds of accidents
are more likely to happen.
Now, I think if AI is used in
an appropriate way, in, in,
in, not even warfare,
but think of just, just
intelligence collection, you know,
let, let's say we're able to,
you know, predict an invasion
of Taiwan or a new movement in Ukraine.
Like, you know, our adversaries
will think twice about,
you know, about conducting
some kind of invasion
or military operation if we know
everything that they're doing.
And so I think superior
intelligence really can de
deter conflict here.
Superior ability to
respond can deter conflict.
I continue to be a
believer in these things.
Anthropic's making headlines
almost on a weekly basis.
Yes. Most notably now
around mythos. Of course,
This is the latest and
greatest Anthropic model,
and it is capable of going
through all the links
of the cyber kill chain
and doing so autonomously.
You said mythos was too
powerful to release to the public.
What surprised you most about it?
I think the thing that
surprised me most about it was the

models had been climbing in their ability
to find vulnerabilities
and importantly turn those
vulnerabilities into exploits,
which people only talk
about the vulnerabilities.
They don't often talk about
turning the vulnerabilities into
exploits, which it, it was quite good at.
So the things that surprised
me are we saw this huge jump.
It was a particularly large jump
and without us really
prompting them at all, some
of the early companies that we gave this
to said things like,
this is a super weapon.
You should have to own
a gun license to use it.
Please don't release this.
Like, the, the demand
to do this was coming from
the companies we gave it to
who were finding so many
critical vulnerabilities
and exploitability around
these critical vulnerabilities
that, you know, they, they were
basically asking us not to,
not to not to release it.
Now to be clear,
'cause things always get
distorted in the world
of social media, the goal isn't
to keep this locked up forever.
We're kind of gradually trying
to open this up to a wider
and wider set of people
and eventually we believe that
we should release mythos to
to, you know, to a general audience,
but with kind of strong cyber safeguards.

Now, a concern is
today's cyber safeguards,
which we did release on Opus 4.7,
which is a good cyber model,
but a substantially weaker one.
These can be jailbroken
and we're a little concerned about some
of the other companies who
think this is a sufficient
defense because yeah, it works sometimes,
but you know, we all know
that these classifiers can
be jailbroken or gone around
and our own testing as well
as frankly our assessment of
the models that other, the defenses
that other companies have
put in place suggests
that these defenses are not strong yet.
And, and that's what we're
waiting for, getting the defenses
to the point where we really
have confidence in them.
There was a lot of pushback on it.
You know, you have researchers
saying they were able
to replicate it using, you know,
cheaper open source models.
Some folks say OpenAI, you know,
has these capabilities already,
you know, what do you say
to folks who say this is a
grand PR play? Yeah, so the,
the claim that could be replicated
with open source models,
that's just incredibly false.
So the idea is mythos looks
across the whole cold base
and finds something.
Some guy went on Twitter
and said, well, if you point
an open source model at exactly

the line of code that mythos finds,
then it finds the same issue.
That, that, that isn't
the, that isn't the prompt,
that isn't the question, right?
Like, that is not, that
is not the same thing.
The ultimate test of this
is like, we go to companies,
we go to open source repos.
We found 271 new
vulnerabilities in Firefox.
We found many thousands
within the private, you know,
companies who haven't fixed them yet
or can't disclose them yet.
Like no one found those 271
vulnerabilities with the previous model.
So like the actual workflow of
what actually works in practice
as opposed to, you know,
okay, I find the exact line
that mythos found, you know,
I found the needle in the haystack.
Something else can now pick up the needle.
But what
about
the folks who say, this
was just good marketing.
You know, we have suffered
enormously commercially from
not releasing this model.
This model has incredibly
accelerated research within
Anthropic and production and next models.
It would do the same in the
outside world if we were to release it.
This has hurt us enormously commercially.
If this helps defenders,
it also helps attackers.
Can we defend anything anymore?
What I would say is that the reason

that we're giving Mythos to defenders
before we give it to attackers
is to patch all the bugs.
I don't know, as the models
get better, there may be more
and more bugs to be found,
but there's only so many,
they're finite, right?
It's like you have this surface
and there's only so many holes in it.
You, you patch all the holes
and the surface becomes
very hard to attack, as well
as the code itself is written
with the powerful models.
So it's, it's then becomes very hard
to find flaws in or break into.
So I think on the other side
of this, hopefully six months
or a year from now, we have
a much more secure internet
ecosystem than we had in the past.
We're trying to get to that world
and we're doing the best
we can to open up mythos
to new cyber defenders.
We've been talking to the
government, we're very respectful
of their recommendations.
They're slowing the pace
at which we open it up
because they're worried about
counterintelligence risk.
I think that's sensible.
I think all serious people here understand
that there's real trade offs here.
We see a lot of sniping
from people on Twitter
and from, you know,
from other AI companies.
You look at what they're saying
and the inconsistency

with what they're doing.
It's not, they're not serious people.
They're not seriously engaging with the,
the serious trade-offs that,
that, that, that we have here.
Look, I have customers calling
me up every day saying,
I want access to Mythos.
I have countries call calling
me up saying, I want access
to Mythos and I have the US government
and my security team
saying, no, wait a minute.
There's risk to it. What?
You know, I'm not saying one
side or the other is right?
I think it's somewhere in between.
Both sides have valid points,
but there's a real challenge here
and we need to face it
together as a society,
not accuse things of
being cheap marketing,
not use cheap marketing to try
and counter position, which some
of the other companies are doing it.
It just, it just all shows
an an incredible lack
of gravitas and maturity.
We need to all face these
mo this moment together.
Have you had to make trade-offs already
that you're not entirely comfortable with?
Throughout the entire history
of Anthropic has been trade-offs, right?
The entire history of Anthropic, right?
Where, you know, in,
in, in, in, in, in, in,
in some ideal world, you
would, you know, you would,
you would prefer to, before
you release the first chatbot,

you know, you could, you
could spend years studying,
you know, every possible thing
that could go wrong with it.
Now, we did delay, we did
delay the initial release
of Claude, but you know,
we did it for a few months.
So what I'm saying is
everything is a trade off.
You know, the, the extreme
ends of the spectrum are, are,
are completely insane.
And so everything is a trade off.
What I would say is that
now that we're in, you know,
what I would describe as a
commercially leading position, I,
I am actually, and Daniela
are actually doing all we can
to move the dial even further towards,
towards being careful.
That's what the Mythos
release was, was about, right?
It's very hard to do something like
that if you're not the leading player.
And so I think you're
gonna see more things,
more things like that.
You know, there's this argument,
why wouldn't the government take you over?
Why would they let a
private company control
technology that's so powerful?
So I actually think that's a very,
that's a very serious question
and I share those concerns.
I don't think the government
should outright take us over,
but I would put it this way.
I would say, just to back up
and describe the situation,

every previous powerful technology
we've seen in history was
either built by the government
or originated with the government.
So nuclear weapons, obviously,
you know, initially built
by the government and pretty much
built by the government after that.
But even like the internet,
GPS, cell phones, all the R&D
was, you know, was done in the labs
and the federal labs and the universities.
AI is the first technology
that's been built in the private sector
and where government has not
really had a serious role
and is coming in late to the game.
I think that's actually a
dangerous and unstable situation.
It is not the situation I would've chosen.
There's not really an
alternative, like, you know,
this technology is possible to build,
our adversaries are building,
it has economic value,
like it's, it's going to get built.
The, the issue is the
government not doing it,
not the private sector doing it.
I think we need to think about
checks and balances on power.
So I think there need to be checks
and balances on the power
of the AI companies, right?
We have this thing, the
long term benefit trust.
What that is, is it's a set
of, basically it's a body
that can appoint the
majority of the board members
and remove the majority
of the board members.

So it basic, it essentially,
if you thread it through,
has the power to fire me.
And what we're looking at
is we're introducing some
elements, you know, nowhere
near all the elements,
but we're introducing a a
little bit of the elements
of like public governance, right?
Where, where it's like, you
know, you're accountable
to someone who just doesn't,
he doesn't just have stock
stock in the company.
So that's, that's very important
and that structure is gonna
continue no matter what happens
to the company. That's on the AI...
And we encourage other companies
to have similar structures
on the government side, I think
we need checks and balances.
You know, there are, there
are efforts in Congress
that have been announced to
enact those red lines, right?
So I really think the, you
know, the legislative branch
and the judicial branch
need to exert themselves
because this technology, I'm
scared of companies having it,
but I'm also scared of
government having it.
And then the companies need to
provide checks on government,
and the government needs to
provide checks on companies.
You know, we need basic
regulation of the technology.
You know, I think we need

to start doing pre-release
testing, required
pre-release testing, testing
and auditing of the models.
You know, it, it's very funny to me
how there's a particular group
of people in the tech world
in Silicon Valley who started,
you know, they, they, they
started with a position
of like even having transparency
around this technology,
even export control, you
know, this is all, you know,
just totally, it'll apocalyptically
destroy our potential
to create the technology.
It'll kill innovation.
And then as soon as they
see the first real danger,
which I've been expecting all
along, there's all this talk
of like nationalization
and the government should just seize it.
Come on folks here, you're,
yo you're yo-yoing from
like the most extreme
anti-regulatory, you
know, you know, if you,
if you look at us the wrong way,
you're destroying the
industry to, you know,
this completely communist,
the government should grab it all.
We, we need a more, we need a more
sensible, moderate approach.
That's the one we've
been favoring all along
because we've, we've understood the
power of this technology.
We're not panicking, we're not denying it.
We see the smooth exponential

and we're responding to it appropriately.
So how was your visit
back to the White House?
You know, we always try to work together
with whoever we can in government.
You know, I, I I said we
have this simple approach,
like we have a set of principles,
we like follow those principles
and we hope that folks on the
other side are reasonable.
And you know, honestly,
the government has taken
Mythos very seriously.
Like we've had good conversations
with Treasury Secretary Bessent
with Chief of Staff, Susie Wiles.
I think they really understand, you know,
the nature of the risks here.
Mythos has, I think, you know, helped them
to feel much more concretely
where these risks are.
So, you know, again, as
with any administration,
there are parts who we
get along with very well
and who understand it.
And you know, there are other parts
that are harder to get along with.
I think that's normal.
That would be the case in
any, in any administration,
and we just try to
navigate it as best we can.
You worked at Baidu
earlier in your career,
big Chinese tech company.
You worked at the Silicon
Valley outpost of it,
and you've been clear
on your views on China.
Strong open source models

are coming out of China
and you have US companies building on them
for free. Is that a threat?
So, you know, one of
the things we've seen
with this technology is that
there's really a premium to
how intelligent the models are.
We very, very rarely see
that people would prefer
to use models with lower intelligence.
Now to be clear, there's
a thriving ecosystem.
There are lots of challenges
and problems that are much
easier than, you know,
the ones we need frontier models for.
But again, it's an exponential, right?
Like it's possible
that like these far from frontier
models have economic value
comparable to what we
saw in 2023 and 2024.
But again, we have this 10X
a year growth and, and,
and so what, what we find is
that what's on the frontier is
always much, much larger than
what is, what is away from the frontier.
I think this is something
that people who are used
to building products in
the previous era don't
quite understand, right?
As someone who's come in who,
you know, hadn't run a company
before, who's like, you know,
has never thought about
the previous product era
particularly to test the social media era,
I feel like an outsider to that world.
And I, I, I feel that

people's instincts are wrong.
They have all these kind
of product heuristics
and I think the 10x per
year model exponential really
breaks that like intelligence
is just such a huge factor
that it outweighs everything else.
And so we're just seeing over
and over again that, you know,
the value is found on the frontier.
Now what I do worry about with some
of these laggard models
is the risks of them
where we have Mythos
class cyber capabilities.
12 months from now,
we'll have much better cyber capabilities,
but the Mythos class cyber
capabilities may just be
available for, for anyone
to, to download. Now,
hopefully we'll have patched
everything before then.
I don't think there's
anything we can do to stop it,
but I, I think it's a serious concern.
Did what you saw by do
shape your views on China?
Not really, no. I worked there for,
I worked there for a year.
You know, I think I probably
learned more about like speech,
speech recognition and,
and you know, all, all, all all of that.
Maybe the only thing that
concerned me was, you know, part
of how we got all the speech
recognition data was, you know,
they're like, they said ominously, oh,
we don't care about privacy in China.
So we have all this, this

speech recognition data.
But I think, I think
aside from my worries,
here are geopolitical, you
know, I think the things
that most worried me about
what happened in China are,
you know, what we, what we
saw happen to the Uyghurs,
what we saw with suppression
of criticism even in the US,
with what happened with Hong Kong, right?
The fact that the CCP could
reach into the US business
network and, you know,
and suppress criticism, that's
an authoritarian state and,
and a high tech authoritarian state.
And when I see how that
combines with AI you,
you really get a, a dystopia
here like 1984 or worse.
And, and my focus is on
trying to prevent that.
And I think we have an
opportunity to prevent that.
I think we have an opportunity for AI
to be a pro-democracy
technology, you know, that kind
of makes people freer that
delivers on the promise
of equal justice for all.
Or it could go the other way. And,
and which way it goes depends on the
actions of the AI companies.
It depends on the actions
of the government,
depends on the actions of all of us.
And so I see us as having
a responsibility here.
There's a moment that people
in your field talk about

where AI gets good
enough to improve itself
and then the improved version
improves itself and so on.
Some of your researchers think
that that moment is close.
How far away is it?
I don't think it's a moment in time.
I think it's a continuous process.
We're already seeing it in
some ways where the AI is able
to suggest architectures for the next AI.
You know, I would say a
year ago, we were seeing 10
to 15% kind
of increase in total factor
productivity due to AI.
Like that's probably up to 20
or 30% now might, you know,
might, it might be doubling like
as with all things we're
on the exponential,
there's no moment where AI improves itself
or runs out of control or becomes unsafe.
What we have is an
accelerating exponential
and at each point on the
exponential we have to assess is,
is this a time to slow down?
Is this a time to, you know,
put more controls on on this technology?
I think more and more of
that is gonna be required.
But I, you know, I think
the Rosetta Stone to all
of this is the smooth exponential.
Again, I think there's an
object lesson in the people
who were against all AI regulation
and then they saw one thing
and they wanted to nationalize.
I think there's an object

lesson in the people
who dismissed the power of AI
and then said, oh my God,
it's improving itself.
It's running outta control.
We have to shut it all down.
Yo-yoing between those extreme
reactions is incredibly
unhelpful as a response
to this technology.
The right response, the wise response is
to say, we're not gonna panic.
Our countermeasures
will smoothly ratchet up
with the power of the technology.
If you see someone having this
kind of crazy yo-yo reaction,
that's a sign that they
were caught by surprise
and that they're not serious.
I understand one of
your favorite books is The
Making of the Atomic Bomb.
That is correct.
Do you see parallels
between yourself and Oppenheimer?
You know, the figure I most identified
with was Leo Szilard,
who, you know, the one
who first basically had the
idea that there could be a kind
of chain reaction.
Look, my view is we're we're
not gonna get through this
with like larger than life personalities
or like figures who try
and be at the center of everything, right?
There needs to be a balance
of power here, right?
There's a lot of powerful
actors who have interests here,
and the only way it's gonna end well

for everyone is if there is
some, there's basically checks
and balances everywhere.
So in some ways I actually see
Oppenheimer as a failure case
as what should not happen.
You've said there's
roughly a 10 to 25% chance
of civilizational collapse.
That is not insignificant.
Is there a scenario where it's something
that Anthropic built, that caused that?
I mean, I certainly hope not.
My view is that, you
know, the, the actions
that we have taken lower
that probability rather
than increasing it, right?
That probability comes
from the, the, you know,
the very straightforward
recipe of the technology,
the existence of many countries
in the world, the existence
of many companies within an
economy and new ones created.
If the void is isn't filled, like
that's a dilemma that we're in.
We are trying to act to
lower that probability.
I think we lower it a lot
more than we raise it.
But, you know, the inherent
property of this technology is
that it's unpredictable.
And so, you know, we
try to build something
and test it a lot before it's released.
And then the models
that are released today are not dangerous,
or at least not, you know,
really I think dangerous

outside of cyber.
And then we try and iterate
and learn from that.
So there's like a zillion defense
mechanisms. You know, half
of what we do within the company is try
and, you know, reduce the
risk as much as we can,
but, you know, it's,
it's never gonna be zero.
I guess what I would say is, you know,
suppose there are a
bunch of like, you know,
airline companies out there
and you're like, well, I'm gonna make an
airline company that's safer.
It can both be the case that, you know,
your airline company is 10
times safer than all the other
airline companies.
But you know, if, if someone comes
and asks you, like, can you guarantee
that your airplane will never crash?
I mean, how could you,
how could you possibly,
But if there was a 25%
chance of an airplane crashing,
you wouldn't get on that plane.
That's right. 25% is too high.
We're trying to make
that probability much,
much lower. That is the goal.
You are building something
incredibly powerful
and stand to gain enormously from it.
Why should we trust you?
My view of this is actually
when any company starts out
and, and particularly,
you know, what we've seen
with the behavior of, of just
Silicon Valley as an entity.

It's, it's thinking over
the last couple years.
I think starting from a
position of distrust, you know,
if you don't know anything about me,
if you know anything about
Anthropic is pretty rational.
I think Silicon Valley has
lost a lot of the world's trust
and kind of has to re-earn it.
And the message, you know,
we're trying to send is
or actually different and,
and that has to be earned in
things that we actually do.
You can agree or disagree, but
we stood up for our values.
The thing with, you
know, Mythos, like it's,
it's really hampered us commercially not
to put this very powerful model out.
And there were a bunch of smaller things
before it, you know,
we, we, we put our money
where our mouth is on, you know, China,
we cut off access to, to models.
We didn't have to do that.
No one told us to do that.
You know, that cost us several
hundred million dollars back
when several hundred
million dollars was a big,
was a significant fraction of our revenue.
You know, the, the delay of Claude two,
like we have a long history of it.
We aren't perfect, we make mistakes.
But you know, what I
would ask is for people
to look at the overall history
and say, if you add up
that overall history,
what is the hypothesis about us

that is most consistent
with that overall history?
People have to decide for themselves,
but I think the hypothesis
that's consistent is we are
genuinely trying to do the right thing.
We're imperfect
organizations are, you know,
always dysfunctional.
We're always trying to, you know, fix them
and make them work better.
Many foot faults, many
things that go wrong,
but at basis we, we have a honest
and earnest picture of
how to do the right thing
and we're trying to
execute on that picture.
We will see you on the other
side of the exponential then
Hopefully.
You always wanted to
be a Hollywood star, right?
I, I, that's one surprising thing
that I didn't understand
about the CEO job is
how often you have to wear makeup.
That was not on my bingo card,
Just a little powder.