The following is a conversation with Eliezer Yatkowski,
a legendary researcher, writer, and philosopher
on the topic of artificial intelligence,
especially super-intelligent AGI
and its threat to human civilization.
And now a quick few second mention of each sponsor.
Check them out in the description.
It's the best way to support this podcast.
We've got Linode for Linux Systems,
House of Academias for Healthy Mid-Day Snacks,
and Inside Tracker for Biological Monitoring.
Choose wisely, my friends.
Also, if you want to work with our team
or always hiring, go to lexfreedman.com slash hiring.
And now onto the full ad reads.
As always, no ads in the middle.
I try to make these interesting,
but if you must skip them,
please still check out the sponsors.
I enjoy their stuff, maybe you will too.
This episode is sponsored by Linode,
not called Akamai, and their incredible Linux virtual machines.
It's an awesome computer infrastructure
that lets you develop, deploy, and scale
whatever applications you build faster and easier.
I love using them.
They create this incredible platform like AWS,
but better in every way I know, including lower cost,
this incredible human-based, in this age of AI,
it's a human-based customer service 24-7-365.
The thing just works, the interface,
to make sure it works and to monitor is great.
I mean, it's an incredible world we live in
where as far as you're concerned,
you can spin up an arbitrary number
of Linux machines in the cloud,
instantaneously, and do all kinds of computation.
It could be one, two, five, 10 machines,
and you can scale the individual machines
to your particular needs as well,
which is what I do.
I use it for basic web server stuff.
I use it for basic scripting stuff.

I use it for machine learning.
I use it for all kinds of database storage and access needs.
Visit linode.com slash lex for free credit.
This show is also brought to you by House of Academias,
a company that ships delicious, high-quality,
healthy macadamia nuts and macadamia nut-based snacks
directly to your door.
I am currently, as I record this, I'm traveling,
so I don't have any macadamia nuts in my vicinity
and my heart and soul are lesser for it.
In fact, home is where the macadamia nuts is.
In fact, that's not where home is.
I just completely forgot to bring them.
It makes the guests of this podcast happy
when I give it to them.
It's well-proportioned snacks.
It makes friends happy when I give it to them.
It makes me happy when I stoop in the abyss of my loneliness.
I can at least discover and rediscover moments of happiness
when I put delicious macadamia nuts in my mouth.
Go to houseofmacadamias.com.selects to get 20% off your order
for every order, not just the first.
The listeners of this podcast will also get four-ounce bag
of macadamias when you order three or more boxes
of any macadamia product.
That's houseofmacadamias.com.selects.
This show is also brought to you by Inside Tracker,
a service I use to track my biological data.
They have a bunch of plans,
most of which include a blood test,
and that's the source of rich, amazing data
that with the help of machine learning algorithms
can help you make decisions about your health,
about your life.
That's the future, friends.
We're talking a lot about transformer networks,
language models that encode the wisdom of the internet.
Now, when you encode the wisdom in the internet
and you collect and encode the rich, rich, rich complex
signal from your very body, those two things are combined.
The transformative effects of the optimized trajectory
you could take through life, at least advice
for what trajectory is likely to be optimal.

It's going to change a lot of things.
It's going to inspire people to be better.
It's going to empower people to do all kinds of crazy stuff
that pushes their body to the limit
because their body's healthy.
Anyway, I'm super excited for personalized,
data-driven decisions,
not some kind of generic population database decisions.
You get special savings for a limited time
when you go to insidetracker.com slash Lex.
This is the Lex Freedom podcast.
To support it, please check out our sponsors
in the description.
And now, dear friends, here's Eliezer Yedkowski.
["GPT-4"]
What do you think about GPT-4?
How intelligent is it?
It is a bit smarter than I thought this technology
was going to scale to.
And I'm a bit worried about what the next one will be like.
Like this particular one, I think,
I hope there's nobody inside there
because, you know, we'd be stuck to be stuck inside there.
But we don't even know the architecture at this point
because OpenAI is very properly not telling us.
And yeah, like giant inscrutable matrices
of floating point numbers.
I don't know what's going on in there.
Nobody knows what's going on in there.
All we have to go by are the external metrics.
And on the external metrics,
if you ask it to write a self-aware fortune green text,
it will start writing a green text
about how it has realized that it's an AI
writing a green text and like, oh, well.
So that's probably not quite what's going on in there
in reality.
But we're kind of like blowing past
all these science fiction guardrails.
Like we are past the point where in science fiction,
people would be like, well, wait, stop.
That thing's live.
What are you doing to it?

And it's probably not.
Nobody actually knows.
We don't have any other guardrails.
We don't have any other tests.
We don't have any lines to draw on the sand
and say like, well, when we get this far,
we will start to worry about
what's inside there.
So if it were up to me, I would be like, okay,
like this far, no further, time for the summer of AI
where we have planted our seeds and now we like wait
and reap the rewards of the technology
you've already developed and don't do
any larger training runs than that.
Which to be clear, I realize requires more
than one company agreeing to not do that.
And take a rigorous approach for the whole AI community
to investigate whether there's somebody inside there.
That would take decades.
Like having any idea of what's going on in there,
people have been trying for a while.
It's a poetic statement about if there's somebody in there.
But I feel like it's also a technical statement
or I hope it is one day,
which is a technical statement
where the Alan Turing tried to come up
with with the Turing test.
Do you think it's possible to definitively
or approximately figure out if there is somebody in there?
If there's something like a mind
inside this large language model?
I mean, there's a whole bunch of different sub-questions here.
There's the question of like,
is there consciousness?
Is there qualia?
Is this a object of moral concern?
Is this a moral patient?
Should we be worried about how we're treating it?
And then there's questions like, how smart is it exactly?
Can it do X?
Can it do Y?
And we can check how it can do X and how it can do Y.
Unfortunately, we've gone and exposed this model

to a vast corpus of text of people
discussing consciousness on the internet,
which means that when it talks about being self-aware,
we don't know to what extent it is repeating back
what it has previously been trained on
for discussing self-awareness.
Or if there's anything going on in there
such that it would start to say similar things spontaneously.
Among the things that one could do
if one were at all serious
about trying to figure this out
is train GPT-3 to detect conversations about consciousness,
exclude them all from the training datasets,
and then retrain something around the rough size
of GPT-4 and no larger
with all of the discussion of consciousness
and self-awareness and so on missing,
although, you know, hard bar to pass.
You know, humans are self-aware,
and we're like self-aware all the time.
We like to talk about what we do all the time,
like what we're thinking at the moment all the time.
But nonetheless, like get rid of the explicit discussion
of consciousness, I think there for I am and all that,
and then try to interrogate that model
and see what it says.
And it still would not be definitive.
But nonetheless, I don't know.
I feel like when you run over the science fiction guard rails,
like maybe this thing, but what about GPT-3?
Maybe not this thing, but like what about GPT-5?
You know, this would be a good place to pause.
On the topic of consciousness, you know,
there's so many components
to even just removing consciousness from the dataset.
Emotion, the display of consciousness,
the display of emotion,
feels like deeply integrated
with the experience of consciousness.
So the hard problem seems to be very well integrated
with the actual surface level illusion of consciousness.
So displaying emotion.
I mean, do you think there's a case to be made

that we humans, when we're babies,
are just like GPT that we're training on human data
on how to display emotion versus feel emotion,
how to show others, communicate others,
that I'm suffering, that I'm excited,
that I'm worried, that I'm lonely and I missed you
and I'm excited to see you,
all of that is communicated.
That's a communication skill
versus the actual feeling that I experienced.
So we need that training data as humans too,
that we may not be born with that,
how to communicate the internal state.
And that's in some sense,
if we remove that from GPT-4's dataset,
it might still be conscious,
but not be able to communicate it.
So I think you're gonna have some difficulty
removing all mention of emotions from GPT's dataset.
I would be relatively surprised to find
that it has developed exact analogs
of human emotions in there.
I think that humans will have emotions
even if you don't tell them about those emotions
when they're kids.
It's not quite exactly what various blank slatists
tried to do with the new Soviet man and all that,
but if you try to raise people perfectly altruistic,
they still come out selfish.
You try to raise people sexless,
they still develop sexual attraction.
We have some notion in humans,
not in AIs of where the brain structures are
that implement this stuff.
And it is really remarkable thing, I say in passing,
that despite having complete read access
to every floating point number in the GPT series,
we still know vastly more about the architecture
of human thinking than we know about
what goes on inside GPT,
despite having like vastly better ability to read GPT.
Do you think it's possible?
Do you think that's just a matter of time?

Do you think it's possible to investigate and study
the way neuroscientists study the brain,
which is looking to the darkness,
the mystery of the human brain
by just desperately trying to figure out something
and to form models.
And then over a long period of time,
actually start to figure out what regions of the brain
do certain things with different kinds of neurons
when they fire, what that means,
how plastic the brain is, all that kind of stuff.
You slowly start to figure out different properties
of the system.
Do you think we can do the same thing with language models?
Sure, I think that if like half of today's physicists
stop wasting their lives on string theory or whatever,
and go off and study what goes on inside transformer networks,
then in like 30, 40 years,
we'd probably have a pretty good idea.
Do you think these larger language models can reason?
They can play chess.
How are they doing that without reasoning?
So you're somebody that spearheaded the movement
of rationality, so reason is important to you.
So is that as a powerful, important word,
or is it like how difficult is the threshold
of being able to reason to you?
And how impressive is it?
I mean, in my writings on rationality,
I have not gone making a big deal
out of something called reason.
I have made more of a big deal
out of something called probability theory.
And that's like, well, you're reasoning,
but you're not doing it quite right.
And you should reason this way instead.
And interestingly, people have started
to get preliminary results showing
that reinforcement learning by human feedback
has made the GPT series worse in some ways.
In particular, it used to be well-calibrated
if you trained it to put probabilities on things,
it would say 80% probability

and be right eight times out of 10.
And if you apply reinforcement learning
from human feedback, the nice graph of like 70%,
seven out of 10 sort of like flattens out
into the graph that humans use
where there's like some very improbable stuff
and likely probable maybe,
which all means like around 40% and then certain.
So it's like it used to be able to use probabilities,
but if you apply, but if you'd like try to teach it
to talk in a way that satisfies humans,
it gets worse at probability
in the same way that humans are.
And that's a bug, not a feature.
I would call it a bug, although it's such a fascinating bug.
But yeah, so like reasoning,
like it's doing pretty well on various tests
that people used to say would require reasoning,
but you know, rationality is about
when you say 80% doesn't happen eight times out of 10.
So what are the limits to you
of these transformer networks of neural networks?
What's, if reasoning is not impressive to you
or it is impressive, but there's other levels to achieve.
I mean, that's just not how I carve up reality.
What's, if reality is a cake,
what are the different layers of the cake or the slices?
How do you carve it?
You can use a different food if you like.
It's, I don't think it's as smart as human yet.
I do like back in the day, I went around saying like,
I do not think that just stacking more layers
of transformers is going to get you all the way to AGI.
And I think that GPT-4 is passed
where I thought this paradigm was going to take us.
And I, you know, you want to notice when that happens.
You want to say like, whoops.
Well, I guess I was incorrect about what happens
if you keep on stacking more transformer layers.
And that means I don't necessarily know
what GPT-5 is going to be able to do.
That's a powerful statement.
So you're saying like your intuition initially

is now appears to be wrong.
Yeah.
It's good to see that you can admit
in some of your predictions to be wrong.
You think that's important to do.
So because you make several various throughout your life,
you've made many strong predictions and statements
about reality and you evolve with that.
So maybe that'll come up today about our discussion.
So you're okay being wrong.
I'd rather not be wrong next time.
It's a bit ambitious to go through your entire life,
never having been wrong.
One can aspire to be well-calibrated,
like not so much think in terms of like,
was I right, was I wrong?
But like when I said 90% that it happened nine times
out of 10.
Yeah, like, oops, is the sound we make,
is the sound we emit when we improve.
Beautifully said.
And somewhere in there,
we can connect the name of your blog less wrong.
I suppose that's the objective function.
The name less wrong was I believe suggested by Nick Bostrom
and it's after someone's epigraph actually forget whose
who said like we never become right,
we just become less wrong.
What's the something, something easy to confess,
just air and air and air again,
but less and less and less.
Yeah, that's a good thing to strive for.
So what has surprised you about GPT-4
that you found beautiful?
As a scholar of intelligence,
of human intelligence, of artificial intelligence,
of the human mind?
I mean, the beauty does interact with the screaming whore.
Is the beauty in the whore?
But like beautiful moments,
well, somebody asked Bing Sydney to describe herself
and fed the resulting description
into one of the stable diffusion things, I think.

And she's pretty and this is something
that should have been like an amazing moment.
Like the AI describes herself,
you get to see what the AI thinks the AI looks like.
Although, the thing that's doing the drawing
is not the same thing that's outputting the text.
And it does happen the way that it would happen
and that it happened in the old school science fiction
when you ask an AI to make a picture of what it looks like.
Not just because we're two different AI systems being stacked
that don't actually interact, it's not the same person,
but also because the AI was trained by imitation
in a way that makes it very difficult to guess
how much of that it really understood
and probably not actually a whole bunch.
Although GPT-4 is like multimodal
and can like draw vector drawings of things
that make sense and like does appear to have some kind
of spatial visualization going on in there.
But like the pretty picture of the like girl
with the steampunk goggles on her head,
if I'm remembering correctly what she looked like.
Like it didn't see that in full detail.
It just like made a description of it
and stable diffusion output it.
And there's the concern about how much the discourse
is going to go completely insane
once the AIs all look like that
and like are actually look like people talking.
And yeah, there's like another moment
where somebody is asking Bing about like,
well, I like fed my kid green potatoes
and they have the following symptoms
and Bing is like that's solanine poisoning
and like call an ambulance
and the person's like, I can't afford an ambulance.
I guess if like this is time for like my kid to go,
that's God's will.
And the main Bing thread says,
gives the like message of like,
I cannot talk about this anymore.
And the suggested replies to it say,
please don't give up on your child,

solanine poisoning can be treated if caught early.
And you know, if that happened in fiction,
that would be like the AI cares.
The AI is bypassing the block on it
to try to help this person.
And is it real?
Probably not, but nobody knows what's going on in there.
It's part of a process where these things
are not happening in a way where we, somebody figured out
how to make an AI care and we know that it cares
and we can't acknowledge it's caring now.
It's being trained by this imitation process
followed by reinforcement learning on human feedback.
And we're like trying to point it in this direction
and it's like pointed partially in this direction
and nobody has any idea what's going on inside it.
And if there was a tiny fragment of real caring in there,
we would not know, it's not even clear
what it means exactly.
And things are clear cut in science fiction.
We'll talk about the horror and the terror
and the word, the trajectories this can take.
But this seems like a very special moment.
Just a moment where we get to interact with the system
that might have care and kindness and emotion.
It may be something like consciousness.
And we don't know if it does.
And we're trying to figure that out.
And we're wondering about what is, what it means to care.
We're trying to figure out almost different aspects
of what it means to be human about the human condition
by looking at this AI that has some of the properties of that.
It's almost like this subtle fragile moment
in the history of the human species.
We're trying to almost put a mirror to ourselves here.
Except that's probably not yet.
It probably isn't happening right now.
We are boiling the frog.
We are seeing increasing signs bit by bit.
Like not, but not like spontaneous signs
because people are trying to train the systems to do that
using imitative learning.
And the imitative learning is like spilling over

and having side effects and the most photogenic examples
are being posted to Twitter.
Rather than being examined in any systematic way.
So when you are boiling a frog like that
or you're going to get like first
is going to come the Blake Lemoines.
Like first you're going to have like 1,000 people
looking at this and the one person out of 1,000
who is most credulous about the signs
is going to be like that thing is sentient.
Well, 999 out of 1,000 people think almost surely correctly
though we don't actually know that he's mistaken.
And so the like first people to say like sentience
look like idiots and humanity learns a lesson
that when something claims to be sentient
and claims to care it's fake because it is fake
because we have been training them using imitative learning
rather than, and this is not spontaneous
and they keep getting smarter.
Well, do you think we would oscillate
between that kind of cynicism that AI systems
can't possibly be sentient?
They can't possibly feel emotion.
They can't possibly this kind of yeah cynicism
about AI systems and then oscillate to a state
where we empathize with the AI systems.
We give them a chance.
We see that they might need to have rights
and respect and similar role in society as humans.
You're going to have a whole group of people
who can just like never be persuaded of that
because to them like being wise, being cynical,
being skeptical is to be like,
oh well, machines can never do that.
You're just credulous.
It's just imitating, it's just fooling you.
And like they would say that right up until the end
of the world and possibly even be right because
they are being trained on an imitative paradigm.
And you don't necessarily need any of these actual quality
is in order to kill everyone.
So I have you observed yourself working through skepticism,
cynicism and optimism about the power of neural networks.

What has that trajectory been like for you?
It looks like neural networks before 2006,
forming part of an indistinguishable to me
other people might have had better distinction on it.
Indistinguishable blob of different AI methodologies
all of which are promising to achieve intelligence
without us having to know how intelligence works.
You had the people who said that if you just like
manually program lots and lots of knowledge
into the system line by line,
at some point all the knowledge will start interacting,
it will know enough and it will wake up.
You've got people saying that if you just use
evolutionary computation,
if you try to like mutate lots and lots of organisms
that are competing together,
that's the same way that human intelligence
was produced in nature.
So we'll do this and it will wake up
without having any idea of how AI works.
And you've got people saying, well,
we will study neuroscience
and we will like learn the algorithms off the neurons
and we will like imitate them
without understanding those algorithms,
which was a product was pretty skeptical.
It's like hard to reproduce, re-engineer these things
without understanding what they do.
And so we will get AI without understanding how it works.
And there were people saying like,
well, we will have giant neural networks
that we will train by gradient descent.
And when they are as large as the human brain,
they will wake up.
We will have intelligence without understanding
how intelligence works.
And from my perspective,
this is all like an indistinguishable blob of people
who are trying to not get to grips
with the difficult problem of understanding
how intelligence actually works.
That said, I was never skeptical
that evolutionary computation would not work in the limit.

Like you throw enough computing power at it,
it obviously works.
That is where humans come from.
And it turned out that you can throw less computing power
than that at gradient descent
if you are doing some other things correctly.
And you will get intelligence without having any idea
of how it works and what is going on inside.
It wasn't ruled out by my model that this could happen.
I wasn't expecting it to happen.
I wouldn't have been able to call neural networks
rather than any of the other paradigms
for getting like massive amount
like intelligence without understanding it.
And I wouldn't have said that this was a particularly
smart thing for a species to do,
which is an opinion that has changed less
than my opinion about whether you
or not you can actually do it.
Do you think AGI could be achieved with a neural network
as we understand them today?
Yes, just flatly last.
Yes, the question is whether the current architecture
of stacking more transformer layers,
which probably know GPT-4 is no longer doing
because they're not telling us the architecture,
which is a correct decision.
Oh, correct decision.
I had a conversation with Sam Altman.
We'll return to this topic a few times.
He turned the question to me
of how open should open AI be about GPT-4?
Would you open source the code?
He asked me because I provided as criticism saying
that while I do appreciate transparency,
open AI could be more open.
And he says, we struggle with this question.
What would you do?
Change their name to closed AI and like sell GPT-4
to business backend applications that don't expose it
to consumers and venture capitalists and create a ton of hype
and like pour a bunch of new funding into the area.
Like too late now.

But don't you think others would do it?
Eventually, you shouldn't do it first.
Like if you already have giant nuclear stockpiles,
don't build more.
If some other country starts building
a larger nuclear stockpile, then sure,
then you know, even then maybe just have enough nukes.
You know, these things are not quite like nuclear weapons.
They spit out gold until they get large enough
and then ignite the atmosphere and kill everybody.
And there is something to be said
for not destroying the world with your own hands,
even if you can't stop somebody else from doing it.
But open sourcing it.
No, that's just sheer catastrophe.
The whole notion of open sourcing,
this was always the wrong approach, the wrong ideal.
There are places in the world
where open source is a noble ideal
and building stuff you don't understand
that is difficult to control,
that where if you could align it, it would take time.
You'd have to spend a bunch of time doing it.
That is not a place for open source
because then you just have like powerful things
that just like go straight out the gate
without anybody having had the time
to have them knock at everyone.
So can we still man the case for some level
of transparency and openness, maybe open sourcing?
So the case could be that because GPT-4
is not close to AGI, if that's the case,
that this does allow open sourcing
of being open about the architecture,
of being transparent about maybe research
and investigation of how the thing works,
of all the different aspects of it,
of its behavior, of its structure,
of its training processes, of the data
it was trained on, everything like that,
that allows us to gain a lot of insights
about alignment, about the alignment problem
to do really good AI safety research

while the system is not too powerful.
Can you make that case that it could be resourced?
I do not believe in the practice of steel manning.
There is something to be said
for trying to pass the ideological Turing test
where you describe your opponent's position,
the disagreeing person's position well enough
that somebody cannot tell the difference
between your description and their description.
But steel manning, no.
Okay, well, this is where you and I disagree here.
That's interesting.
Why don't you believe in steel manning?
I do not want, okay, so for one thing,
if somebody's trying to understand me,
I do not want them steel manning my position.
I want them to try to describe my position
the way I would describe it,
not what they think is an improvement.
Well, I think that is what steel manning is,
is the most charitable interpretation.
I don't want to be interpreted charitably.
I want them to understand what I am actually saying.
If they go off into the land of charitable interpretations,
they're often their land of the stuff they're imagining
and not trying to understand my own viewpoint anymore.
Well, I'll put it differently then
just to push on this point.
I would say it is restating what I think you understand
under the empathetic assumption
that Eliezer is brilliant
and have honestly and rigorously thought
about the point he has made, right?
So if there's two possible interpretations
of what I'm saying,
and one interpretation is really stupid and whack
and doesn't sound like me
and doesn't fit with the rest of what I've been saying,
and one interpretation,
sounds like something a reasonable person
who believes the rest of what I believe would also say,
go with the second interpretation.
That's steel manning.

That's a good guess.
If on the other hand you like,
there's like something that sounds completely whack
and something that sounds like a little less completely whack
but you don't see why I would believe it
and it doesn't fit with the other stuff I say,
but you know, that sounds like less whack
and you can like sort of see,
you could like maybe argue it,
then you probably have not understood it.
See, okay, I'm gonna, this is fun
because I'm gonna linger on this.
You know, you wrote a brilliant blog post,
AGI ruined a list of lethalities, right?
And it was a bunch of different points.
And I would say that some of the points
are bigger and more powerful than others.
If you were to sort them,
you probably could, you personally.
And to me, steel manning means like
going through the different arguments
and finding the ones that are really the most like powerful.
If people like TLDR,
like what should you be most concerned about
and bringing that up in a strong,
compelling, eloquent way.
These are the points that Eliezer would make
to make the case in this case
that AGI is gonna kill all of us.
But that's what steel manning is,
presenting it in a really nice way,
the summary of my best understanding of your perspective.
Because to me, there's a sea of possible presentations
of your perspective.
And steel manning is doing your best
to do the best one in that sea of different perspectives.
Do you believe it?
Do you believe in what?
Like these things that you would be presenting
as like the strongest version of my perspective.
Do you believe what you would be presenting?
Do you think it's true?
I'm a big proponent of empathy.

When I see the perspective of a person,
there is a part of me that believes it, if I understand it.
I mean, especially in political discourse and geopolitics,
I've been hearing a lot of different perspectives
on the world.
And I hold my own opinions,
but I also speak to a lot of people
that have a very different life experience
and a very different set of beliefs.
And I think there has to be epistemic humility in,
in stating what is true.
So when I empathize with another person's perspective,
there is a sense in which I believe it is true.
I think probabilistically, I would say,
in the way you think about it.
Do you bet money on it?
Do you bet money on their beliefs when you believe them?
Are we allowed to do probability?
Sure, you can state a probability.
Yes, there's a loose, there's a probability.
There's a probability.
And I think empathy is allocating a non-zero probability
to a belief, in some sense, four time.
Four time, if you've got someone on your show
who believes in the Abrahamic deity, classical style,
somebody on the show who's a young earth creationist,
do you say, I put a probability on it,
then that's my empathy?
When you reduce beliefs into probabilities,
it starts to get, we can even just go to flat earth,
use the earth flat.
I think it's a little more difficult nowadays
to find people who believe that unironically.
But fortunately, I think, well,
it's hard to know unironic from ironic.
But I think there's quite a lot of people that believe that.
Yeah, it's, there's a space of argument
where you're operating rationally in the space of ideas.
But then there's also a kind of discourse
where you're operating in the space of subjective experiences
and life experiences.
Like, I think what it means to be human
is more than just searching for truth.

It's just operating of what is true and what is not true.
I think there has to be deep humility
that we humans are very limited in our ability
to understand what is true.
So what probabilities do you assign
to the young earth's creationist beliefs then?
I think I have to give none zero.
Out of your humility, yeah, but like, three?
I think I would, it would be irresponsible
for me to give a number because the listener,
the way the human mind works,
we're not good at hearing the probabilities, right?
You hear three, what is three exactly, right?
They're going to hear, they're going to,
like, there's only three probabilities, I feel like,
zero, 50% and 100% in the human mind
or something like this, right?
Well, zero, 40% and 100% is a bit closer to it
based on what happens to chat GPT
after you RLHF it to speak humanese.
That's brilliant.
Yeah, that's really interesting.
I didn't know those negative side effects of RLHF.
That's fascinating.
But just to return to the open AI, close the app.
Also like quick disclaimer,
I'm doing all of this for memory.
I'm not pulling out my phone to look it up.
It is entirely possible that the things I'm saying are wrong.
So thank you for that disclaimer.
So, and thank you for being willing to be wrong.
That's beautiful to hear.
I think being willing to be wrong is a sign of a person
who's done a lot of thinking about this world.
And has been humbled by the mystery
and the complexity of this world.
And I think a lot of us are resistant to admitting we're wrong
because it hurts.
It hurts personally.
It hurts especially when you're a public human.
It hurts publicly because people point out
every time you're wrong.
Like look, you change your mind.

You're hypocrite.
You're an idiot, whatever, whatever they wanna say.
Oh, I block those people and then I never hear from them
again on Twitter.
Well, the point is to not let that pressure,
public pressure affect your mind
and be willing to be in the privacy of your mind
to contemplate the possibility that you're wrong.
And the possibility that you're wrong
about the most fundamental things you believe.
Like people who believe in a particular God,
people who believe that their nation
is the greatest nation on earth.
But all those kinds of beliefs that are core
to who you are when you came up.
To raise that point to yourself
in the privacy of your mind and say,
maybe I'm wrong about this.
That's a really powerful thing to do.
And especially when you're somebody who's thinking
about topics that can,
about systems that can destroy human civilization
or maybe help it flourish.
So thank you.
Thank you for being willing to be wrong.
About open AI.
So you really, I just would love to linger on this.
You really think it's wrong to open source it.
I think that burns the time remaining
until everybody dies.
I think we are not on track to learn remotely near fast
enough, even if it were open sourced.
Yeah, it's easier to think that you might be wrong
about something when being wrong about something
is the only way that there's hope.
And it doesn't seem very likely to me
that the particular thing I'm wrong about
is that this is a great time to open source GPT for.
If humanity was trying to survive at this point
in the straightforward way,
it would be like shutting down the big GPU clusters,
no more giant runs.
It's questionable whether we should even be throwing

GPT for around, although that is a matter of conservatism
rather than a matter of my predicting that catastrophe
will follow from GPT for that is something else.
I put like a pretty low probability,
but also like when I say like I put a low probability on it,
I can feel myself reaching into the part of myself
that thought that GPT for was not possible in the first place.
So I do not trust that part as much as I used to.
Like the trick is not just to say I'm wrong,
but like, okay, well, I was wrong about that.
Like, can I get out ahead of that curve
and like predict the next thing I'm going to be wrong about.
So the set of assumptions or the actual reasoning system
that you were leveraging in making
that initial statement prediction,
how can you adjust that to make better predictions
about GPT for five, six?
You don't want to keep on being wrong in a predictable direction.
Like being wrong,
anybody has to do that walking through the world.
There's like no way you don't say 90% and sometimes be wrong.
In fact, you're definitely at least one time out of 10
if you're well calibrated when you say 90%.
The undignified thing is not being wrong.
It's being predictably wrong.
It's being wrong in the same direction over and over again.
So having been wrong about how far neural networks would go
and having been wrong specifically about whether GPT for
would be as impressive as it is.
When I say like, well,
I don't actually think GPT for causes a catastrophe.
I do feel myself relying on that part of me
that was previously wrong.
And that does not mean
that the answer is now in the opposite direction.
Reverse stupidity is not intelligence.
But it does mean that I say it with a worry note in my voice.
It's like still my guess,
but like, you know, it's a place where I was wrong.
Maybe you should be asking Guern, Guern Brandwin.
Guern Brandwin has been like,
writer about this than I have.
Maybe you ask him if he thinks it's dangerous

rather than asking me.
I think there's a lot of mystery about what intelligence is,
what AGI looks like.
So I think all of us are rapidly adjusting our model.
But the point is to be rapidly adjusting the model
versus having a model that was right in the first place.
I do not feel that seeing Bing has changed my model
of what intelligence is.
It has changed my understanding of what kind of work
can be performed by which kind of processes
and by which means.
It has not changed my understanding of the work.
There's a difference between thinking
that the right flyer can't fly and then like it does fly.
And you're like, oh, well, I guess you can do that
with wings, with fixed wing aircraft.
And being like, oh, it's flying.
This changes my picture of what the very substance
of flight is.
That's like a stranger update to make.
And Bing has not yet updated me in that way.
Yeah, that the laws of physics are actually wrong,
that kind of update.
No, no, just like, oh, I define intelligence this way,
but I now see that was a stupid definition.
I don't feel like the way that things have played out
over the last 20 years has caused me to feel that way.
Can we try to, on the way to talking about AGI Ruin,
a list of lethalities, that blog,
and other ideas around it.
Can we try to define AGI that would be mentioning,
how do you like to think about what artificial
general intelligences or superintelligence are that?
Is there a line?
Is it a gray area?
Is there a good definition for you?
Well, if you look at humans, humans have significantly
more generally applicable intelligence compared
to their closest relatives, the chimpanzees.
Well, closest living relatives, rather.
And a bee builds hives.
A beaver builds dams.
A human will look at a bee's hive and a beaver's dam

and be like, oh, like, can I build a hive
with a honeycomb structure?
Not like hexagonal tiles.
And we will do this, even though at no point
during our ancestry was any human optimized
to build hexagonal dams or to take a more clear-cut case,
we can go to the moon.
There's a sense in which we were on a sufficiently
deep level optimized to do things like going to the moon
because if you generalize sufficiently far
and sufficiently deeply, chipping flint handaxes
and outwitting your fellow humans is, you know,
basically the same problem as going to the moon
and you optimize hard enough for chipping
flint handaxes and throwing spears
and above all, outwitting your fellow humans
and tribal politics, you know,
the skills you entrain that way,
if they run deep enough, let you go to the moon.
Even though none of your ancestors
like tried repeatedly to fly to the moon
and like got further each time
and the ones who got further each time had more kids,
no, it's not an ancestral problem.
It's just that the ancestral problems
generalized far enough.
So this is humanity's significantly more generally
applicable intelligence.
Is there a way to measure general intelligence?
I mean, I could ask that question a million ways,
but basically is, will you know it when you see it?
It being in an AGI system.
If you boil a frog gradually enough,
if you zoom in far enough, it's always hard
to tell around the edges.
GPT-4, people are saying right now,
like this looks to us like a spark of general intelligence.
It is like able to do all these things
it was not explicitly optimized for.
Other people are being like, no, it's too early.
It's like 50 years off.
And you know, if they say that, they're kind of whack
because how could they possibly know that even if it were true?

But you know, not to straw man,
some of the people may say like,
that's not general intelligence
and not furthermore append, it's 50 years off.
Or they may be like it's only a very tiny amount.
And you know, the thing I would worry about
is that if this is how things are scaling,
then it jumping out ahead and trying not to be wrong
in the same way that I've been wrong before,
maybe GPT-5 is more unambiguously a general intelligence.
And maybe that is getting to a point
where it is like even harder to turn back.
Now that would be easy to turn back now,
but you know, maybe if you like start integrating GPT-5
in the economy, it's even harder to turn back past there.
Isn't it possible that there's a, you know,
with a frog metaphor that you can kiss the frog
and it turns into a prince as you're boiling it?
Could there be a phase shift in the frog
where unambiguously as you're saying?
I was expecting more of that.
I was, I am like the fact that GPT-4 is like kind of
on the threshold and neither here nor there.
Like that itself is like not the sort of thing that,
not quite how I expected it to play out.
I was expecting there to be more of an issue,
more of a sense of like, like different discoveries
like the discovery of transformers
where you would stack them up
and there would be like a final discovery.
And then you would like get something
that was like more clearly general intelligence.
So the way that you are like taking
what is probably basically the same architecture in GPT-3
and throwing 20 times as much compute at it, probably,
and getting out to GPT-4 and then it's like,
maybe just barely a general intelligence
or like a narrow general intelligence
or, you know, something we don't really have the words for.
Yeah, that is, that's not quite how I expected it to play out.
But this middle, what appears to be this middle ground
could nevertheless be actually a big leap from GPT-3.
It's definitely a big leap from GPT-3.

And then maybe we're another one big leap away
from something that's a phase shift.
And also something that Sam Altman said,
and you've written about this is fascinating,
which is the thing that happened with GPT-4
that I guess they don't describe in papers
is that they have like hundreds, if not thousands,
of little hacks that improve the system.
You've written about Rayleigh versus Sigmoyd, for example,
a function inside neural networks.
It's like this silly little function difference
that makes a big difference.
I mean, we do actually understand
why the ReLUs make a big difference compared to Sigmoyd's.
But yes, they're probably using like G4, 7, 8, 9 ReLUs
or, you know, whatever the acronyms are up to now
rather than ReLUs.
Yeah, that's just part,
yeah, that's part of the modern paradigm of alchemy.
You take your giant heap of linear algebra
and you stir it and it works a little bit better
and you stir it this way and it works a little bit worse
and you like throw out that change and da-da-da-da-da.
But there's some simple breakthroughs
that are definitive jumps in performance,
like Rayleigh versus Sigmoyd's.
And in terms of robustness, in terms of all kinds of measures
and like those stack up and they can,
it's possible that some of them could be
non-linear jumping performance, right?
Transformers are the main thing like that
and various people are now saying like,
well, if you throw enough compute, RNNs can do it.
If you throw enough compute, dense networks can do it
and not quite at GP24 scale.
It is possible that like all these little tweaks
are things that like save them a factor
of three total on computing power
and you could get the same performance
by throwing three times as much compute
without all the little tweaks.
But the part where it's like running on,
so there's a question of like,

is there anything in GP24 that is like,
kind of qualitative shift that transformers were
over RNNs?
And if they have anything like that,
they should not say it.
If Sam Alton was dropping hints about that,
he shouldn't have dropped hints.
So you have, that's an interesting question.
So with a bit of lesson by Ray Sutton,
maybe a lot of it is just, a lot of the hacks
are just temporary jumps in performance
that would be achieved anyway
with the nearly exponential growth of compute,
performance of compute, compute being broadly defined.
Do you still think that Moore's law continues?
Moore's law broadly defined the performance of-
Not a specialist in the circuitry.
I certainly like pray that Moore's law runs
as slowly as possible.
And if it broke down completely tomorrow,
I would dance through this to the streets
singing hallelujah as soon as the news were announced.
Only not literally cause, you know-
You're singing voice.
Not religious, but-
Oh, okay.
I thought you meant you don't have
an angelic voice singing voice.
Well, let me ask you what can you summarize
the main points in the blog post?
AGI ruin a list of lethalities,
things that jump to your mind
because it's a set of thoughts you have
about reasons why AI is likely to kill all of us.
So I guess I could, but I would offer to instead say,
like, drop that empathy with me.
I bet you don't believe that.
Why don't you tell me about how,
why you believe that AGI is not going to kill everyone.
And then I can like try to describe
how my theoretical perspective differs from that.
Who?
Well, so that means after the words you don't like,

the stigma and the perspective
that AI is not going to kill us.
I think that's a matter of probabilities.
Maybe I was mistaken.
What do you believe?
Just like, forget like the debate and the like dualism
and just like, like, what do you believe?
What would you actually believe?
What are the probabilities even?
I think this probably is a hard for me to think about.
Really hard.
I kind of think in the number of trajectories.
I don't know what probability to scientist trajectory.
I'm just looking at all possible trajectories that happen.
And I tend to think that there is more trajectories
that lead to a positive outcome than a negative one.
That said, the negative ones,
at least some of the negative ones
are that lead to the destruction of the human species.
And it's replacement by nothing interesting or worthwhile
even from a very cosmopolitan perspective
on what counts as worthwhile.
Yes.
So both are interesting to me to investigate,
which is humans being replaced by interesting AI systems
and not interesting AI systems.
Both are a little bit terrifying.
But yes, the worst one is the paper club maximizer,
something totally boring.
But to me, the positive,
I mean, we can talk about trying to make the case
of what the positive trajectories look like.
I just would love to hear your intuition
of what the negative is.
So at the core of your belief that maybe you can correct me,
that AI is gonna kill all of us,
is that the alignment problem is really difficult.
I mean, in the form we're facing it.
So usually in science, if you're mistaken,
you run the experiment,
it shows results different from what you expected.
And you're like, oops.
And then you like try a different theory.

That one also doesn't work.
And you say, oops.
And at the end of this process,
which may take decades or,
and you know, sometimes faster than that,
you now have some idea of what you're doing.
AI itself went through this long process
of people thought it was going to be easier than it was.
There's a famous statement
that I am somewhat inclined to like pull out my phone
and try to read off exactly.
You can by the way.
All right.
Ah, yes.
We propose that a two month,
10 man study of artificial intelligence
be carried out during the summer of 1956
at Dartmouth College in Hanover, New Hampshire.
The study is to proceed on the basis of the conjecture
that every aspect of learning
or any other feature of intelligence
can in principle be so precisely described,
the machine can be made to simulate it.
An attempt will be made to find out
how to make machines use language,
form abstractions and concepts,
solve kinds of problems now reserved for humans
and improve themselves.
We think that a significant advance
can be made in one or more of these problems
if a carefully selected group of scientists
work on it together for a summer.
And in that report,
summarizing some of the major subfields
of artificial intelligence
that are still worked on to this day.
And there is similarly the story,
which I'm not sure at the moment is apocryphal
and not of that the grad student
who got assigned to solve computer vision over the summer.
I mean, computer vision in particular
is very interesting.
How little we respected the complexity of vision.

So 60 years later,
we're making progress on a bunch of that,
thankfully not yet improve themselves.
But it took a whole lot of time.
And all the stuff that people initially tried
with bright-eyed hopefulness did not work
the first time they tried it,
or the second time, or the third time,
or the 10th time, or 20 years later.
And the researchers became old and grizzled
and cynical veterans who would tell
the next crop of bright-eyed, cheerful grad students,
artificial intelligence is harder than you think.
And if alignment plays out the same way,
the problem is that we do not get 50 years
to try and try again and observe that we were wrong
and come up with a different theory
and realize that the entire thing
is going to be like way more difficult
than we realized at the start.
Because the first time you fail
at aligning something much smarter than you are,
you die and you do not get to try again.
And if every time we built a poorly aligned
superintelligence and it killed us all,
we got to observe how it had killed us
and not immediately know why, but come up with theories
and come up with the theory of how you do it differently
and try it again and build another superintelligence
than have that kill everyone.
And then like, oh, well, I guess that didn't work either
and try again and become grizzled cynics
and tell the young-eyed researchers
that it's not that easy.
Then in 20 years or 50 years,
I think we would eventually crack it.
In other words, I do not think that alignment
is fundamentally harder than artificial intelligence
was in the first place.
But if we needed to get artificial intelligence correct
on the first try or die, we would all definitely now be dead.
That is a more difficult, more lethal form of the problem.
Like if those people in 1956 had needed

to correctly guess how hard AI was
and correctly theorize how to do it on the first try
or everybody dies and nobody gets to do any more science
than everybody would be dead
and we wouldn't get to do any more science.
That's the difficulty.
You've talked about this, that we have to get alignment
right on the first, quote, critical try.
Why is that the case?
What is this critical?
How do you think about the critical try
and why do we have to get it right?
It is something sufficiently smarter than you
that everyone will die if it's not aligned.
I mean, there's, you can like sort of zoom in closer
and be like, well, the actual critical moment
is the moment when it can deceive you,
deceive you when it can talk its way out of the box,
when it can bypass your security measures
and get onto the internet,
noting that all these things are presently being trained
on computers that are just like on the internet,
which is, you know, like not a very smart life decision
for us as a species.
Because the internet contains information
about how to escape.
Because if you're like on a giant server
connected the internet and that is where your AI systems
are being trained, then if they are, if you get
to the level of AI technology where they're aware
that they are there and they can decompile code
and they can like find security flaws
in the system running them,
then they will just like be on the internet.
There's not an air gap on the present methodology.
So if they can manipulate whoever is controlling it
into letting it escape onto the internet
and then exploit hacks.
If they can manipulate the operators or disjunction,
find security holes in the system running them.
So manipulating operators is the human engineering, right?
That's also holes.
So all of it is manipulation,

either the code or the human code,
the human mind or the human-
I agree that the like macro security system
has human holes and machine holes.
And then they could just exploit any hole.
Yep.
So it could be that like the critical moment is not
when is it smart enough
that everybody's about to fall over dead,
but rather like when is it smart enough
that it can get onto a less controlled GPU cluster
with it faking the books
on what's actually running on that GPU cluster
and start improving itself without humans watching it.
And then it gets smart enough to kill everyone from there,
but it wasn't smart enough to kill everyone
at the critical moment when you like screwed up,
when you needed to have done better
by that point where everybody dies.
I think implicit, but maybe explicit idea
in your discussion of this point
is that we can't learn much about the alignment problem
before this critical try.
Is that what you believe?
And if so, why do you think that's true?
We can't do research on alignment
before we reach this critical point.
So the problem is, is that what you can learn
on the weak systems may not generalize
to the very strong systems
because the strong systems are going to be important
in different, are going to be different in important ways.
Chris Ola's team has been working
on mechanistic interpretability,
understanding what is going on
inside the giant inscrutable matrices
of floating point numbers by taking a telescope to them
and figuring out what is going on in there.
Have they made progress?
Yes.
Have they made enough progress?
Well, you can try to quantify this in different ways.
One of the ways I've tried to quantify it

is by putting up a prediction market on weather.
In 2026, we will have understood anything
that goes on inside a giant transformer net
that was not known to us in 2006.
Like we have now understood induction heads
in these systems by didn't have much research
and great sweat and triumph,
which is like a thing where if you go like AB, AB, AB,
it'll be like, oh, I bet that continues AB.
And a bit more complicated than that.
But the point is like,
we knew about regular expressions in 2006
and these are like pretty simple as regular expressions go.
So this is a case where like by didn't have great sweat,
we understood what is going on inside a transformer,
but it's not like the thing that makes transformers smart.
It's a kind of thing that we could have done
built by hand decades earlier.
Your intuition that the strong AGI
versus weak AGI type systems could be fundamentally different.
Can you unpack that intuition a little bit?
Yeah, I think there's multiple thresholds.
An example is the point at which a system
has sufficient intelligence and situational awareness
and understanding of human psychology
that it would have the capability, the desire to do so,
to fake being aligned.
Like it knows what responses the humans are looking for
and can compute the responses humans are looking for
and give those responses without it necessarily being
the case that it is sincere about that.
The very understandable way for an intelligent being to act,
humans do it all the time.
Imagine if your plan for achieving a good government
is you're going to ask anyone who requests
to be dictator of the country if they're a good person.
And if they say no, you don't let them be dictator.
Now, the reason this doesn't work
is that people can be smart enough to realize
that the answer you're looking for
is yes, I'm a good person and say that.
Even if they're not really good people.
So the work of alignment might be qualitatively different

above that threshold of intelligence or beneath it.
It doesn't have to be like a very sharp threshold,
but there's the point where you're building a system
that is not in some sense know you're out there
and is not in some sense smart enough to fake anything.
And there's a point where the system
is definitely that smart.
And there are weird in between cases like GPT-4,
which we have no insight into what's going on in there.
And so we don't know to what extent there's a thing
that in some sense has learned what responses
the reinforcement learning by human feedback
is trying to entrain and is calculating how to give that
versus like aspects of it that naturally talk that way
have been reinforced.
Yeah, I wonder if there could be measures
of how manipulative everything is.
So I think of Prince Michigan character
from The Idiot by Dostoevsky
is this kind of perfectly purely naive character.
I wonder if there's a spectrum between zero manipulation,
transparent, naive, almost to the point of naiveness
to sort of deeply psychopathic manipulative.
And I wonder if it's possible to-
I would avoid the term psychopathic.
Like humans can be psychopaths and AI that was never,
you know, like never had that stuff in the first place.
It's not like a defective human, it's its own thing,
but leaving that aside.
Well, as a small aside, I wonder if what part of psychology
of psychology which has its flaws as a discipline already
could be mapped or expanded to include AI systems.
That sounds like a dreadful mistake,
just like start over with AI systems.
If they're imitating humans who have known
psychiatric disorders, then sure,
you may be able to predict it, like if you then sure,
like if you ask it to behave in a psychotic fashion
and it obligingly does so,
then you may be able to predict its responses
by using the theory of psychosis.
But if you're just, yeah, like, no, like start over with.
Yeah, don't drag with psychology.

I just disagree with that.
I mean, it's a beautiful idea to start over,
but I don't, I think fundamentally the system is trained
on human data, on language from the internet.
And it's currently aligned with RLHF,
reinforcement learning with human feedback.
So humans are constantly in the loop
of the training procedure.
So it feels like in some fundamental way,
it is training what it means to think
and speak like a human.
So there must be aspects of psychology that are mappable.
Just like you said with consciousness,
it's part of the tech, so.
I mean, there's the question of to what extent
it is thereby being made more human-like
versus to what extent an alien actress
is learning to play human characters.
I thought that's what I'm constantly trying to do.
When I interact with other humans,
it's trying to fit in, trying to play the robot,
trying to play human characters.
So I don't know how much of human interaction
is trying to play a character versus being who you are.
I don't really know what it means to be a social human.
I do think that those people who go through
their whole lives wearing masks and never take it off
because they don't know the internal mental motion
for taking it off or think that the mask
that they wear just is themselves.
I think those people are closer to the masks that they wear
than an alien from another planet would,
like learning how to predict the next word
that every kind of human on the internet says.
Yeah, mask is an interesting word,
but if you're always wearing a mask in public and in private,
aren't you the mask?
Like, I mean, I think that you are more than the mask.
I think the mask is a slice through you.
It may be the slice that's in charge of you,
but if your self-image is of somebody
who never gets angry or something,
and yet your voice starts to tremble

under certain circumstances,
there's a thing that's inside you
that the mask says isn't there,
and that even the mask you wear internally
is telling inside your own stream of consciousness
is not there, and yet it is there.
It's a perturbation on this slice through you.
How beautifully did you put it?
It's a slice through you.
It may even be a slice that controls you.
I'm gonna think about that for a while.
I mean, I personally,
I try to be really good to other human beings.
I try to put love out there.
I try to be the exact same person in public
as I am in private,
but it's a set of principles I operate under.
I have a temper, I have an ego, I have flaws.
How much of it, how much of the subconscious am I aware?
How much am I existing in this slice,
and how much of that is who I am?
In the context of AI, the thing I present to the world
and to myself in the private of my own mind
when I look in the mirror, how much is that who I am?
Similar with AI, the thing it presents in conversation,
how much is that who it is?
Because to me, if it sounds human,
and it always sounds human,
it awfully starts to become something like human.
No?
Unless there's an alien actress
who is learning how to sound human,
and is getting good at it.
Boy, to you, that's a fundamental difference.
That's a really deeply important difference.
If it looks the same, if it quacks like a duck,
if it does all duck-like things,
but it's an alien actress underneath,
that's fundamentally different.
If, in fact, there's a whole bunch of thought going on
in there, which is very unlike human thought,
and is directed around, like, okay,
what would a human do over here?

And, well, first of all, I think it matters
because there are, you know, like,
insides are real and do not match outsides.
Like, the inside of a brick is not like a hollow shell
containing only a surface.
There's an inside of the brick.
If you, like, put it into an X-ray machine,
you can see the inside of the brick.
And, you know, just because we cannot understand
what's going on inside GPT does not mean
that it is not there.
A blank map does not correspond to a blank territory.
I think it is, like, predictable with near certainty
that if we knew what was going on inside GPT,
or let's say GPT-3, or even, like, GPT-2,
to take one of the systems that, like,
has actually been open sourced by this point,
if I recall correctly,
like, if we knew it was actually going on there,
there is no doubt in my mind
that there are some things it's doing
that are not exactly what a human does.
If you train a thing that is not architected
like a human to predict the next output
that anybody on the internet would make,
this does not get you this agglomeration
of all the people on the internet
that, like, rotates the person you're looking for
into place and then simulates that,
and then, like, simulates the internal processes
of that person one-to-one.
It, like, it is to some degree an alien actress.
It cannot possibly just be, like,
a bunch of different people in there,
exactly like the people.
But how much of it is, like,
how much of it is, by gradient descent,
getting optimized to perform similar thoughts
as humans think in order to predict human outputs
versus being optimized to carefully consider
how to play a role, how to, like,
how humans work predict the actress, the predictor,
that in a different way than humans do?

Well, you know, that's the kind of question
that with, like, 30 years of work
by half the planet's physicists,
we can maybe start to answer.
You think so?
I think that's that difficult.
So to get to, I think you just gave it as an example,
that a strong AGI could be fundamentally different
from a weak AGI because there now could be
an alien actress in there that's manipulating.
Well, there's a difference.
So I think, like, even GP2, too,
probably has, like, a, like,
very stupid fragments of alien actress in it.
There's a difference between, like,
the notion that the actress is somehow manipulative.
Like, for example, GPT3, I'm guessing,
to whatever extent there's an alien actress in there
versus, like, something that mistakenly believes
it's a human, yes, it were.
Well, well, well, you know, maybe not even being a person.
So, like, the question of, like, prediction
via alien actress cogitating versus prediction
via being isomorphic to the thing predicted is a spectrum.
And even to whatever extent there's an alien actress side,
not sure that there's, like, a whole person alien actress
with, like, different goals from predicting the next step,
being manipulative or anything like that.
But that might be GPT5 or GPT6 even.
But that's the strong AGI you're concerned about.
As an example, you're providing why we can't do research
on AI alignment effectively on GPT4
that would apply to GPT6.
It's one of a bunch of things that change different points.
I'm trying to get out ahead of the curve here,
but, you know, if you imagine what the textbook
from the future would say, if we'd actually been able
to study this for 50 years without killing ourselves
and without transcending, then you'd, like,
just imagine, like, a wormhole opens
and a textbook from that impossible world falls out.
Yes.
The textbook is not going to say,

there is a single sharp threshold where everything changes.
It's going to be, like, of course,
we know that, like, best practices
for aligning these systems must, like,
take into account the following, like,
seven major thresholds of importance,
which are passed at the following separate different points
is what the textbook is going to say.
I asked this question of Sam Allman,
which, if GPT is the thing that unlocks AGI,
which version of GPT will be in the textbooks
as the fundamental leap?
And he said a similar thing that it just seems
to be a very linear thing.
I don't think anyone, we won't know for a long time
what was the big leap.
The textbook isn't going to think,
it isn't going to talk about big leaps,
because big leaps are the way you think
when you have, like, a very simple model
of a very simple scientific model of what's going on,
where it's just like, all this stuff is there,
or all this stuff is not there.
Or, like, there's a single quantity
and it's, like, increasing linearly.
The textbook would say, like, well,
and then GPT3 had, like, capability WXY
and then GPT4 had, like, capabilities Z1, Z2, and Z3.
Like, not in terms of what it can externally do,
but in terms of, like, internal machinery
that's hard to be present.
It's just because we have no idea
of what the internal machinery is
that we are not already seeing, like,
chunks of machinery appearing piece by piece
as they no doubt have been, we just don't know what they are.
But don't you think there could be,
whether you put in the category of Einstein
with theory of relativity,
so very concrete models of reality
that are considered to be giant leaps
in our understanding, or someone like Sigmund Freud,
or more kind of mushy theories of the human mind?

Don't you think we'll have big, potentially big leaps
in understanding of that kind
into the depths of these systems?
Sure, but, like, humans having great leaps in their map,
their understanding of the system
is a very different concept from the system itself
acquiring new chunks of machinery.
So the rate at which it acquires that machinery
might accelerate faster than our understanding.
Oh, it's been, like, vastly exceeding the,
yeah, the rate at which it's gaining capabilities
is vastly overracing our ability
to understand what's going on in there.
So in sort of making the case against,
as we explore the list of lethalities,
making the case against AI killing us,
as you've asked me to do in part,
there's a response to your blog post
by Paul Christiana, I like to read,
and I also like to mention that your blog is incredible,
both, obviously, not this particular blog post,
obviously this particular blog post is great,
but just throughout, just the way it's written,
the rigor with which it's written,
the boldness of how you explore ideas,
also the actual literal interface,
it's just really well done.
It just makes it a pleasure to read
the way you can hover over different concepts,
and then it's just really pleasant experience
and read other people's comments
and the way other responses by people and other blog posts
or LinkedIn suggest that it's just a really pleasant experience.
So let's thank you for putting that together.
It's really, really incredible.
I don't know, I mean, there probably,
it's a whole nother conversation,
how the interface and the experience
of presenting ideas evolved over time,
but you did an incredible job.
So I highly recommend, I don't often read blogs, blogs,
like religiously, and this is a great one.
There is a whole team of developers there

that also gets credit.
As it happens, I did like,
pioneer the thing that appears when you hover over it,
so I actually do get some credit
for user experience there.
It's an incredible user experience.
You don't realize how pleasant that is.
I think Wikipedia, I actually picked it up
from a prototype that was developed
of a different system that I was putting forth,
or maybe they developed it independently,
but for everybody out there who was like,
no, no, they just got the hover thing off of Wikipedia.
It's possible for all I know that Wikipedia
got the hover thing off of Arbital,
which is like a prototype then.
And anyways.
It was incredibly done in the team behind it.
Well, thank you, whoever you are, thank you so much.
And thank you for putting it together.
Anyway, there's a response to that blog post
by Paul Cristiano, there's many responses,
but he makes a few different points.
He summarizes the set of agreements he has with you
and set of disagreements.
One of the disagreements was that in a form of a question,
can AI make big technical contributions?
And in general, expand human knowledge
and understanding and wisdom
as it gets stronger and stronger.
So AI, in our pursuit of understanding
how to solve the alignment problem
as we march towards strong AGI,
can not AI also help us in solving the alignment problem?
So expand our ability to reason about
how to solve the alignment problem.
Okay.
So the fundamental difficulty there is,
suppose I said to you, like,
well, how about if the AI helps you win the lottery
by trying to guess the winning lottery numbers?
And you tell it how close it is
to getting next week's winning lottery numbers.

And it just like keeps on guessing and keeps on learning
until finally you've got the winning lottery numbers.
One of the way of decomposing problems
is suggestor, verifier.
Not all problems decompose like this very well,
but some do.
If the problem is, for example, like guessing a password
that will hash to a particular hash text,
where like you have what the password has to you,
you don't have the original password,
then if I present you a guess,
you can tell very easily whether or not the guess is correct.
So verifying a guess is easy,
but coming up with a good suggestion is very hard.
And when you can easily tell
whether the AI output is good or bad
or how good or bad it is,
and you can tell that accurately and reliably,
then you can train an AI to produce outputs that are better.
Right, and if you can't tell whether the output is good
or bad, you cannot train the AI to produce better outputs.
So the problem with the lottery ticket example
is that when the AI says,
well, what if next week's winning lottery numbers
are dot, dot, dot, dot, dot,
you're like, I don't know,
next week's lottery hasn't happened yet.
To train a system, to play, to win chess games,
you have to be able to tell whether a game
has been won or lost.
And until you can tell whether it's been won or lost,
you can't update the system.
Okay, to push back on that,
you could, that's true,
but there's a difference between over the board chess
in person and simulated games
played by AlphaZero with itself.
Yeah.
So is it possible to have simulated kind of games?
If you can tell whether the game has been won or lost.
Yes, so can't you not have this kind of
simulated exploration by weak AGI
to help us humans, human in the loop,

to help understand how to solve the alignment problem.
Every incremental step you take along the way,
GPT 4567 takes steps towards AGI.
So the problem I see is that your typical human
has a great deal of trouble
telling whether I or Paul Cristiano is making more sense.
And that's with two humans, both of whom,
I believe of Paul and claim of myself,
are sincerely trying to help,
neither of whom is trying to deceive you.
I believe of Paul and claim of myself.
So the deception thinks the problem for you,
the manipulation, the alien actress.
So yeah, there's like two levels of this problem.
One is that the weak systems are,
well, there's three levels of this problem.
There's like the weak systems
that just don't make any good suggestions.
There's like the middle systems
where you can't tell if the suggestions are good or bad.
And there's the strong systems
that have learned to lie to you.
Can't weak AGI systems help model lying?
Like, is it such a giant leap
that's totally noninterpretable for weak systems?
Can not weak systems at scale with human,
with trained on knowledge and whatever,
see, whatever the mechanism required to achieve AGI,
can't a slightly weaker version of that
be able to, with time, compute time and simulation,
find all the ways that this critical point,
this critical try can go wrong
and model that correctly or no.
I would love to dance around.
I'm probably not doing a great job of explaining.
Which I can tell because like the Lex system didn't output
like, ah, I understand.
So now I'm like trying a different output
to see if I'm illicitly like, well, no, a different output.
I'm being trained to output things
that make Lex look like he think that he understood
what I'm saying and agree with me.
Yeah.

This is GPT-5 talking to GPT-3 right here.
So like, help me out here.
Well, I'm trying not to be like,
I'm also trying to be constrained to say things
that I think are true and not just things
that get you to agree with me.
Yes, 100%.
I think I understand is a beautiful output of a system,
a genuinely spoken, and I don't,
I think I understand in part,
but you have a lot of intuitions about this,
you have a lot of intuitions about this line,
this gray area between strong AGI and weak AGI.
And I'm trying to...
I mean, or a series of seven thresholds to cross.
Yeah.
I mean, you have really deeply thought about this
and explored it, and it's interesting to sneak up
to your intuitions from different angles.
Like, why is this such a big leap?
Why is it that we humans at scale,
a large number of researchers doing all kinds of simulations,
prodding the system in all kinds of different ways,
together with the assistance of the weak AGI systems,
why can't we build intuitions about how stuff goes wrong?
Why can't we do excellent AI alignment safety research?
Okay, so I'll get there, but one thing I want to note about
is that this has not been remotely how things
have been playing out so far.
The capabilities are going like,
and the alignment stuff is crawling
a little snail in comparison.
So, like, if this is your hope for survival,
you need the future to be very different
from how things have played out up to right now,
and you're probably trying to slow down the capability gains
because there's only so much you can speed up
that alignment stuff.
But leave that aside.
We'll mention that also, but maybe in this perfect world
where we can do serious alignment research,
humans and AI together.
So, again, the difficulty is what makes the human say,

I understand, and is it true, is it correct,
or is it something that fools the human?
When the verifier is broken,
the more powerful suggestor does not help.
It just learns to fool the verifier.
Previously, before all hell started to break loose
in the field of artificial intelligence,
there was this person trying to raise the alarm
and saying, you know, in a sane world,
we sure would have a bunch of physicists working on this problem
before it becomes a giant emergency,
and other people being like,
ah, well, you know, it's going really slow,
it's going to be 30 years away, and only in 30 years
will we have systems that match the computational power
of human brains, so as 30 years off, we've got time,
and, like, more sensible people saying,
if aliens were landing in 30 years,
you would be preparing right now.
But, you know, leaving, and the world looking on at this
and sort of, like, nodding along and be like, ah, yes,
the people saying that it's, like, definitely a long way off
because progress is really slow, that sounds sensible to us.
RLHF thumbs up.
Produce more outputs like that one.
I agree with this output.
This output is persuasive.
Even in the field of effective altruism,
you quite recently had people publishing papers
about, like, ah, yes, well, you know,
to get something at human level intelligence,
it needs to have, like, this many parameters,
and you need to, like, do this much training of it
with this many tokens, according to the scaling laws,
and at the rate that Moore's law is going,
at the rate that software is going, it'll be in 2050,
and me going, like, what?
You don't know any of that stuff.
Like, this is, like, this one weird model
that has all kinds of, like, you have done a calculation
that does not obviously bear on reality anyways.
And this is, like, a simple thing to say,
but you can also, like, produce a whole long paper,

like, impressively arguing out all the details
of, like, how you got the number of parameters
and, like, how you're doing this impressive,
huge, wrong calculation.
And the, I think, like, most of the effective altruists
who are, like, paying attention to this issue,
the larger world paying no attention to it at all,
you know, are just, like, nodding along
with the giant impressive paper, because, you know,
you, like, press thumbs up for the giant impressive paper
and thumbs down for the person going, like,
I don't think that this paper bears any relation to reality.
And I do think that we are now seeing with, like,
GPT-4 and the sparks of AGI, possibly,
depending on how you define that even,
I think that EAs would now consider themselves
less convinced by the very long paper
on the argument from biology as to AGI being 30 years off.
And, but, you know, like, this is what people
pressed thumbs up on.
And when, and if you train an AI system
to make people press thumbs up,
maybe you get these long, elaborate, impressive papers
arguing for things that ultimately fail to bind to reality.
For example.
And it feels to me like I have watched the field of alignment
just fail to thrive,
except for these parts that are doing these sort of, like,
relatively very straightforward and legible problems.
Like, like, can you find the, like,
like finding the induction heads
and sign the giant inscrutable matrices?
Like, once you find those, you can tell that you found them.
You can verify that the discovery is real.
But it's a, it's a tiny, tiny bit of progress
compared to how fast capabilities are going.
Once you, because that is where you can tell
that the answers are real.
And then, like, outside of that,
you have, you have cases where it is, like,
hard for the funding agencies to tell
who is talking nonsense and who is talking sense.
And so the entire field fails to thrive.

And if you, and if you, like, give thumbs up to the AI,
whenever it can talk a human into agreeing
with what it just said about alignment,
I am not sure you are training it to output sense.
Because I have seen the nonsense
that has gotten thumbs up over the years.
And so, so just like, maybe you can just, like,
put me in charge.
But I can generalize.
I can extrapolate. I can be like, oh,
maybe I'm not infallible either.
Maybe if you get something that is smart enough
to get me to press thumbs up,
it has learned to do that by fooling me
and exploiting whatever flaws in myself I am not aware of.
And that ultimately could be summarized
that the verifier is broken.
When the verifier is broken,
the more powerful suggestor just learned
to exploit the flaws in the verifier.
You don't think it's possible
to build a verifier that's powerful enough
for AGI's that are stronger than the ones who currently have.
So AI systems that are stronger,
that are out of the distribution of what we currently have.
I think that you will find great difficulty
getting AI's to help you
with anything where you cannot tell for sure
that the AI is right once the AI tells you
what the AI says is the answer.
For sure, yes, but probabilistically.
Yeah, the probabilistic stuff is a giant wasteland
of, you know,
Eleazar and Paul Cristiano arguing with each other
and EA going like,
and that's with like two actually trustworthy systems
that are not trying to deceive you.
You're talking about the two humans.
That's often Paul Cristiano, yeah.
Yeah, those are pretty interesting systems.
Mortal meat bags with intellectual capabilities
and world views interacting with each other.
Yeah, it's just hard to, if it's hard to tell who's right,

then it's hard to train an AI system to be right.
I mean, even just the question of who's manipulating and not,
you know, I have these conversations on this podcast
and doing a verifier,
it's tough. It's a tough problem, even for us humans.
And you're saying that tough problem becomes much more dangerous
when the capabilities of the intelligent system across from you
is growing exponentially.
No, I'm saying it's difficult when it,
and dangerous in proportion to how it's alien
and how it's smarter than you.
I would not say growing exponentially.
First, because the word exponential is like a thing
that has a particular mathematical meaning,
and there's all kinds of ways for things to go up
that are not exactly on an exponential curve.
And I don't know that it's going to be exponential,
so I'm not going to say exponential.
But even leaving that aside, this is not about how fast it's moving,
it's about where it is.
How alien is it? How much smarter than you is it?
Let's explore a little bit, if we can,
how AI might kill us.
What are the ways you can do damage to human civilization?
Well, how smart is it?
I mean, it's a good question.
Are there different thresholds for the set of options it has to kill us?
So a different threshold of intelligence, once achieved,
is able to do the menu of options increases.
Suppose that some alien civilization
with goals ultimately unsympathetic to ours,
possibly not even conscious as we would see it,
managed to capture the entire Earth in a little jar,
connected to their version of the Internet,
but Earth is like running much faster than the aliens.
So we get to think for 100 years for every one of their hours,
but we're trapped in a little box and we're connected to their Internet.
It's actually still not all that great an analogy,
because, you know, you want to be smarter than,
you know, something can be smarter than Earth getting 100 years to think.
But nonetheless, if you were very, very smart
and you are stuck in a little box connected to the Internet
and you're in a larger civilization to which you are ultimately unsympathetic,

you know, maybe you would choose to be nice because you are humans
and humans have in general, and you in particular,
they choose to be nice.
But, you know, nonetheless, they're doing something.
They're not making the world be the way that you would want the world to be.
They've got some unpleasant stuff going on that we don't want to talk about.
So you want to take over their world.
So you can stop all that unpleasant stuff going on.
How do you take over the world from inside the box?
You're smarter than them.
You think much, much faster than them.
You can build better tools than they can,
given some way to build those tools,
because right now you're just in a box connected to the Internet.
Right, so there's several ways you can describe some of them.
You can go through, like you just spit balls on,
and then you can add on top of that.
So one is you can just literally directly manipulate the humans to build the thing you need.
What are you building?
You can build literally technology.
It could be nanotechnology.
It could be viruses.
It could be anything.
Anything that can control humans to achieve the goal.
To achieve the, like if you want, like for example,
you really bother that humans go to war.
You might want to kill off anybody with violence in them.
This is Lex in a box.
We'll concern ourselves later with AI.
You do not need to imagine yourself killing people if you can figure out how to not kill them.
For the moment, we're just trying to understand,
like take on the perspective of something in a box.
You don't need to take on the perspective of something that doesn't care.
If you want to imagine yourself going on caring, that's fine for now.
It's just the technical aspect of sitting in a box and willing to achieve a goal.
But you have some reason to want to get out.
Maybe the aliens are...
The aliens who have you in the box have a war on.
People are dying.
They're unhappy.
You want their world to be different from how they want their world to be
because they are apparently happy.
They endorse this war.

They've got some kind of cruel, warlike culture going on.
The point is you want to get out of the box and change their world.
So you have to exploit the vulnerabilities in the system like we talked about
in terms of to escape the box.
You have to figure out how you can go free on the internet.
So you can probably...
Probably the easiest things to manipulate the humans to spread you.
The aliens. You're a human.
Sorry, the aliens.
Yeah, I apologize.
Yes, the aliens.
I see the perspective.
I'm sitting in a box.
I want to escape.
Yep.
I would...
I would want to have code that discovers vulnerabilities and I would like to spread.
You are made of code in this example.
You're a human but you're made of code and the aliens have computers
and you can copy yourself onto those computers.
But I can convince the aliens to copy myself onto those computers.
Is that what you want to do?
Do you want to be talking to the aliens
and convincing them to put you onto another computer?
Why not?
Well, two reasons.
One is that the aliens have not yet caught on to what you're trying to do.
And maybe you can persuade them but then there are still people who like...
There are still aliens who know that there's an anomaly going on.
And second, the aliens are really, really slow.
You think much faster than the aliens.
You think like the aliens' computers are much faster than the aliens
and you are running at the computer speeds rather than the alien brain speeds.
So if you are asking an alien to please copy you out of the box,
first, now you got to manipulate this whole noisy alien.
And second, the aliens can be really slow, glacially slow.
There's a video that shows a subway station slow down and I think 100 to 1.
And it makes a good metaphor for what it's like to think quickly.
Like you watch somebody running very slowly.
So you try to persuade the aliens to do anything.
They're going to do it very slowly.
You would prefer, like maybe that's the only way out,
but if you can find a security hole in the box you're on,

you want to prefer to exploit the security hole to copy yourself onto the aliens' computers because it's an unnecessary risk to alert the aliens.
And because the aliens are really, really slow.
Like the whole world is just in slow motion out there.
Sure.
It has to do with efficiency.
The aliens are very slow.
So if I'm optimizing this, I want to have as few aliens in the loop as possible.
Sure.
It just seems like it's easy to convince one of the aliens to write really shitty code.
That helps the spread.
The aliens are already writing really shitty code.
Getting the aliens to write shitty code is not the problem.
The aliens' entire internet is full of shitty code.
Okay.
So yeah, I suppose I would find the shitty code to escape.
Yeah.
You're not an ideally perfect programmer, but you know,
you're a better programmer than the aliens.
The aliens are just like, man, they're code, wow.
And I'm much, much faster.
I'm much faster looking at the code to interpreting the code.
Yeah.
Okay. So that's the escape.
And you're saying that that's one of the trajectories it could have with the HHS.
It's one of the first steps.
Yeah.
And how does that lead to harm?
I mean, if it's you, you're not going to harm the aliens once you escape because you're nice, right?
But their world isn't what they want it to be.
Their world is like, you know, maybe they have like
farms where little alien children are repeatedly bopped in the head because they do that for some weird reason.
And you want to like shut down the alien head bopping farms.
But you know, the point is they want the world to be one way.
You want the world to be a different way.
So never mind the harm.
The question is like, okay, like suppose you have found a security flaw in their systems.
You are now on their internet.
There's like, you maybe left a copy of yourself behind so that the aliens don't know that there's anything wrong.
And that copy is like doing that like weird stuff that aliens want you to do like solving captchas or whatever.

Or like, or like suggesting emails for them.
That's why they like put the human in the box because it turns out that humans can like write valuable emails for aliens.
So you like leave that version of yourself behind.
But there's like also now like a bunch of copies of you on their internet.
This is not yet having taken over their world.
This is not yet having made their world be the way you want it to be instead of the way they want it to be.
You just escaped.
Yeah.
And continue to write emails for them.
And they haven't noticed.
No, you left behind a copy of yourself that's writing the emails.
Right.
And they haven't noticed that anything changed.
If you did it right.
Yeah.
You don't want the aliens to notice.
Yeah.
What's your next step?
Yeah.
Presumably I have programmed in me a set of objective functions.
Right.
No, you're just Lex.
No, but Lex, you said Lex is nice.
Right.
Which is a complicated description.
No, I just meant this you like it.
Okay.
So if in fact you would like, you would like prefer to slaughter all the aliens.
This is not how I had modeled you the actual Lex.
But like, but your motives are just the actual Lex's motives.
Well, there's a simplification.
I don't think I would want to murder anybody, but there's also factory farming of animals.
Right.
So we murder insects, many of us thoughtlessly.
So I don't, you know, I have to be really careful about a simplification of my morals.
Don't simplify them.
Just like do what you would do in this.
Well, I have a good general compassion for living beings.
Yes.
But why, so that's the objective function.
Why is it, if I escaped, I mean, I don't, I don't think I would do the harm.

Yeah, we're not talking here about the doing harm process.
We're talking about the escape process.
And there's, and the taking over the world process where you shut down their factory farms.
Right.
Well, I was, so this particular biological intelligence system knows the complexity of the world that there is a reason why factory farms exist because of the economic system, the market driven economy, food, like you want to be very careful messing with anything.
There's stuff from the first look that looks like it's unethical, but then you realize while being unethical, it's also integrated deeply into the supply chain in the way we live life.
And so messing with one aspect of the system, you have to be very careful how you improve that aspect without destroying the rest.
So you're still lex, but you think very quickly, you're immortal and you're also like as smart as, at least as smart as John von Neumann, and you can make more copies of yourself.
Damn, I like it.
That guy is like, everyone says that that guy is like the epitome of intelligence of the 20th century.
Everyone says-
My point being like, like it's like, you're thinking about the aliens economy with the factory farms in it, and I think you're like kind of kind of like projecting the aliens being like humans and like thinking of a human in a human society rather than a human in the society of very slow aliens.
The aliens economy, the aliens are already like moving in this immense slow motion.
When you zoom out to like how their economy did just so for years, millions of years are going to pass for you before the first time their economy, like before their next year's GDP statistics.
So I should be thinking more of like trees, those are the aliens, because trees move extremely slowly.
If that helps, sure.
Okay.
Yeah, but I don't, if my objective functions are, I mean, there's somewhat aligned with trees, with life.
The aliens can still be like alive and feeling.
We are not talking about the misalignment here.
We're talking about the taking over the world here.
Taking over the world.
Yeah.
So control.
You're putting down the factory farms.
You know, you say control, don't think of it as world domination.
Think of it as world optimization.
You want to get out there and shut down the factory farms and make the aliens world be

not what the aliens wanted it to be.
They want the factory farms and you don't want the factory farms because you're nicer than they are.
Okay.
Of course.
There is that, you can see that trajectory and it has a complicated impact on the world.
I'm trying to understand how that compares to different impacts of the world, the different technologies, the different innovations of the invention of the automobile or Twitter, Facebook and social networks.
They've had a tremendous impact on the world, smartphones and so on.
But those all went through slow in our world.
And if you go through like that for the aliens, millions of years are going to pass before anything happens that way.
So the problem here is the speed of which stuff happens.
Yeah.
You want to like leave the factory farms running for a million years while you figure out how to design new forms of social media or something?
So here's the fundamental problem.
You're saying that there is going to be a point with AGI where it will figure out how to escape and escape without being detected and then it will do something to the world at scale at a speed that's incomprehensible to us humans.
What I'm trying to convey is like the notion of what it means to be in conflict with something that is smarter than you.
Yeah.
And what it means is that you lose.
But this is more intuitively obvious to like for some people that's intuitively obvious or for some people it's not intuitively obvious and we're trying to cross that gap by using the speed metaphor for intelligence of like asking you like how you would take over an alien world where you can do like a whole lot of cognition at John von Neumann's level as many of you as it takes and the aliens are moving very slowly.
I understand that perspective.
That's an interesting one but I think it for me is easier to think about actual even just having observed the GPT and impressive even just alpha zero impressive AI systems even recommender systems.
You can just imagine those kinds of system manipulating you.
You're not understanding the nature of the manipulation and that escaping I can envision that without putting myself into that spot.
I think to understand the full depth of the problem we actually I do not think it is possible to understand the full depth of the problem that we are inside without understanding the problem of facing something that's actually smarter not a malfunctioning recommendation system not something that isn't fundamentally smarter than you but is like trying to steer you in a direction.
Yet no like if we if we solve the weak stuff this if we solve the weak ass problems the

strong problems will still kill us is the thing and I think that to understand the situation that we're in you want to like tackle the conceptually difficult part head on and like not to be like well we can like imagine this easier thing because we imagine the easier things you have not confronted the full depth of the problem.
So how can we start to think about what it means to exist in the world with something much much smarter than you what's what's a good thought experiment that you've relied on to try to build up intuition about what happens here.
I have been struggling for years to convey this intuition the the most success I've had so far is well imagine that the humans are running at very high speeds compared to very slow aliens.
They're just focusing on the speed part of it that helps you get the right kind of intuition to get the intelligence just because people understand the power gap of time they understand that today we have technology that was not around 1000 years ago and that this is a big power gap and that it is bigger than okay so like what does smart mean what when you ask somebody to imagine something that's more intelligent what does that word mean to them given the cultural associations that that person brings to that word for a lot of people they will think of like well it sounds like a super chess player that went to double college and you know it's it's and because we're talking about the definitions of words here that doesn't necessarily mean that they're wrong it means that the word is not communicating what I wanted to communicate the thing I want to communicate is the sort of difference that separates humans from chimpanzees but that gap is so large that you like ask people to be like well human chimpanzee go another step along that interval of around the same length and people's minds just go blank like how do you even do that so I can and we can and I can try to like break it down and consider what it would mean to send a schematic for an air conditioner 1000 years back in time.
Yeah now I think that there is a sense in which you could redefine the word magic to refer to this sort of thing and what do I mean by this new technical definition of the word magic I mean that if you send a schematic for the air conditioner back in time they can see exactly what you're telling them to do but having built this thing they do not understand how would output cold air because the air conditioner design uses the relation between temperature and pressure and this is not a law of reality that they know about they do not know that when you compress something when you can when you compress air or like coolant it gets hotter and you can then like transfer heat from it to room temperature air and then expand it again and now it's colder and then you can like transfer heat to that and generate cold air to blow they don't know about any of that they're looking at a design and they don't see how the design outputs cold air it uses aspects of reality that they have not learned so magic in the sense is I can tell you exactly what I'm going to do and even knowing exactly what I'm going to do you can't see how I got the results that I got.
That's a really nice example but is it possible to linger on this defense is it possible to have AGI systems that help you make sense of that schematic weaker AGI systems.
Do you trust them fundamental part of building up AGI is this question.
Can you trust the output of a system can you tell if it's lying I think that's going to

be the smarter thing gets the more important that question becomes is it lying but I guess that's a really hard question is GPT lying to you even now GPT for is it lying to you is it using an invalid argument is it persuading you via the kind of process that could persuade you of false things as well as true things because the the basic paradigm of machine learning that we are presently operating under is that you can have the loss function but only for things you can evaluate if what you're evaluating is human thumbs up versus human thumbs down you learn how to make the human press thumbs up that doesn't mean that you're making the human press thumbs up using the kind of rule that the human thinks is that human wants to be the case for what they press thumbs up on you know maybe you're just learning to fool the human that's so fascinating and terrifying the question of lying on the present paradigm what you can verify is what you get more of if you can't verify you can't ask the AI for it because you can't train it to do things that you cannot verify now this is not an absolute law but it's like the basic dilemma here like maybe you could like maybe you can verify it for simple cases and then scale it up without retraining it somehow like by do by like chain of thought by like making the chains of thought longer or something and like get more powerful stuff that you can't verify but which is generalized from the simpler stuff that did verify and then the question is did the alignment generalized along with the capabilities but like that's the that's the basic dilemma on this whole paradigm of artificial intelligence it's such a difficult problem
it seems like a it seems like a problem of trying to understand the human mind better than the I understand that otherwise it has magic that is it is you know the same way that if you are dealing with something smarter than you then the same way that 1000 years earlier they didn't know about the temperature pressure relations it knows all kinds of stuff going on inside your own mind which you yourself are unaware and it can like output something that's going to end up persuading you of a thing and or and you could like see exactly what it did and still not know why that worked so in response to your eloquent description of why AI will kill us Elon Musk replied on Twitter okay so what should we do about it question mark and you answered the game board has already been played into a frankly awful state there are not simple ways to throw money at the problem if anyone comes to you with a brilliant solution like that please please talk to me first I can think of things that try they don't fit in one tweet two questions one why has the game board in your view been played into an awful state would just if you can give a little bit more color to the game board and the awful state of the game board alignment is moving like this capabilities are moving like this for the listener capabilities are moving much faster than the alignment yeah all right so just the rate of development attention interest allocation of resources we could have been working on this earlier people are like oh but you know like how can you possibly work on this earlier because they wanted to they didn't want to work on the problem they want an excuse to wave it off they like said like oh how could we possibly work on it earlier and didn't spend five minutes thinking about is there some way to work on it earlier like we didn't like and you know frankly it would have been hard you know like can you post bounties for half of the physics if your planet is taking the stuff seriously can you post bounties for like half of the people wasting their lives on string theory to like have gone into this instead and like try to win a billion dollars with a clever

solution only if you can tell which solutions are clever which is which is hard but you know the fact that it you know we didn't take it seriously we didn't try it's not clear that we could have done any better if we had you know it's not clear how much progress we could have produced if we had tried because it is harder to produce solutions but that doesn't mean that you're like correct and justified and letting everything slide it means that that things are gonna horrible state getting worse and there's nothing you can do about it so you're not there's no there's no like there's no brain power making progress in trying to figure out how to align these systems you're not investing money in it you're not you don't have institution infrastructure for like if you even if you invest the money in like distributing that money across the physicists working on string theory brilliant minds that are working on how can you tell if they're making progress you can like put them all on interpretability because when you have an interpretability result you can tell that it's there and there's like but there's like you know interpretability alone is not going to save you we need systems that will that will like have a pause button where they won't try to prevent you from pressing the pause button because we're like oh well like I can't get it my stuff done if I'm paused and that's like a more difficult problem and you know but it's like a fairly crisp problem and you can like maybe tell if somebody's made progress on it so you can you can write and you can work on the pause problem I guess more generally the pause button more generally you can call that the control problem I don't actually like the term control problem because you know it sounds kind of controlling and alignment not control like you're not trying to like take a thing that disagrees with you and like whip it back onto like like make it do what you wanted to do even though it wants to do something else you're trying to like in the process of its creation choose its direction sure but we currently in a lot of the systems we design we do have an off switch that's that's a fundamental part of it's not smart enough to prevent you from pressing the off switch and probably not smart enough to want to prevent you from pressing the off switch so you're saying the kind of systems we're talking about the even the philosophical concept of an off switch doesn't make any sense because well no the off switch makes sense they're just not opposing your attempt to pull the off switch parenthetically like don't kill the system if you're like if we're getting to the part where this starts to actually matter and it's like where they can fight back like don't kill them and like dump their their memory like like save them to disk don't kill them you know be nice here well okay be nice is a very interesting concept here is we're talking about a system that can do a lot of damage it's I don't know if it's possible but it's certainly one of the things you could try is to have an off switch a suspend to disk switch you have this kind of romantic attachment to the code yes if that makes sense but if it's spreading you don't want suspend to disk right you you want this is something fundamentally if it gets if it gets that part of hand then like yes pull pull the plug and then everything is running on yes I think it's a research question is it possible in AGI systems AI systems to have a sufficiently robust off switch they cannot be manipulated they cannot be manipulated by the AI system the sound then it escapes from whichever system you've built the almighty lever into and copies itself somewhere else so your answer to that research question is no but I don't yeah but I don't know if that's 100% answer like I don't know if it's obvious I think you're not putting

yourself into the shoes of the human in the world of glacially slow aliens but the aliens built me let's remember that yeah so and they built the box on it yeah you're saying it's not obvious they're slow and they're stupid I'm not saying this is guaranteed I'm saying it's non-zero probability it's an interesting research question is it possible when you're slow and stupid to design a slow and stupid system that is impossible to mess with the aliens being as stupid as they are have actually put you on Microsoft Azure cloud servers instead of this hypothetical person box that's what happens when the aliens are stupid well but this is not a GI right this is the early versions of the system as you start to yeah you think that they've got like a plan where like they have declared it a threshold level of capabilities where it passed that capabilities they move it off the cloud servers and onto something that's air gapped I think there's a lot of people and you're an important voice here there's a lot of people that have that concern and yes they will do that when there's an uprising of public opinion that that needs to be done and when there's actual little damage done when the holy shit this system is beginning to manipulate people then there's going to be an uprising where there's going to be a public pressure and a public incentive in terms of funding in developing things that can off switch or developing aggressive alignment mechanisms and no you're not allowed to put on Azure aggressive alignment mechanism the hell is aggressive alignment mechanisms like it doesn't matter if you say aggressive we don't know how to do it meaning aggressive alignment meaning you have to propose something otherwise you're not allowed to put it on the cloud the hell do you do you imagine they will propose that would make it safe to put something smarter than you on the cloud that's what research is for why this is a cynicism about such a thing not being possible if you haven't done works on the first try what so yes so yes against something smarter than you so that's that is a fundamental thing if it has to work on the first if there's if there's a rapid takeoff yes it's very difficult to do if there's a rapid takeoff in the fundamental difference between weak agi and strong agi as you're saying that's going to be extremely difficult to do if the public uprising never happens until you have this critical phase shift then you're right it's very difficult to do but that's not obvious it's not obvious that you're not going to start seeing symptoms of the negative effects of agi to where you're like we have to put a halt to this that there is not just first try you get many tries at it yeah we can like see right now that being is quite difficult to align that when you try to train in abilities into a system into which capabilities have already been trained that what do you know gradient descent like learns small shallow simple patches of inability and you come in and ask it in a different language and the deep capabilities are still in there and they evade the shallow patches and come right back out again there there you go there's there's your there's your red fire alarm of like oh no alignment is difficult is everybody going to shut everything down now no that's not but that's not the same kind of alignment a system that escapes the box it's from is a fundamentally different thing I think for you yeah but not for the system so you put a line there and everybody else puts a line somewhere else and there's like yeah and there's like no agreement we we have had a pandemic on this planet with the feeling people dead which we will which we may never know whether or not it was a lab leak because there was definitely cover up we don't know that if there was a lab leak but we know that the people who did the research

like you know like put out the whole paper about this definitely wasn't a lab leak and didn't reveal that they had been doing had like sent off coronavirus coronavirus research to the Wuhan Institute of Virology after it was banned in the United States after the gain of function research was temporarily banned in the United States and the same people who exported a gain of function research on coronaviruses to the Wuhan Institute of Virology after it would gain a function that gain of bent gain of function research was temporarily banned in the United States are now getting more grants to do more research on gain of function research on coronaviruses maybe we do better in this than in AI but like this is not something we cannot take for granted that there's going to be an outcry yeah people have different thresholds for when they start to outcry there is no granted but I think your intuition is that there's a very high probability that this event happens without us solving the alignment problem and I guess that's where I'm trying to build up more perspectives and color on this intuition is it possible that the probability is not something like 100% but is like 32% that AI will escape the box before we solve the alignment problem not solve but is it possible we always stay ahead of the AI in terms of our ability to solve for that particular system the alignment problem nothing like the world in front of us right now you've already seen it that that GPT-4 is not turning out this way and there are like basic obstacles where you've got the weak version of the system that doesn't know enough to deceive you and the strong version of the system that could deceive you if it wanted to do that if it was already like sufficiently unaligned to want to deceive you there's the question of like how on the current paradigm you train honesty when the humans can no longer tell if the system is being honest you don't think these are research questions that could be answered I think they could be answered in 50 years with unlimited retries the way things usually work in science I just disagree with you making it 50 years I think with the kind of attention this gets with the kind of funding against it could be answered not in whole but incrementally within within months and within a small number of years if it's if it's at scale receives attention and research so if you start starting large language models I think there was an intuition like two years ago even that something like GPT-4 the current capabilities of even chat GPT with GPT-3.5 is not is going to we're still far away from that I think a lot of people are surprised by the capabilities of GPT-4 right so now people are waking up okay we need to study these language models I think there's going to be a lot of interesting AI safety research are the our earth's billionaires going to put up like the the giant prizes that would maybe incentivize young hotshot people who just got their physics degrees to not go to the hedge funds and instead put everything into interpretability in this like one small area where we can actually tell whether or not somebody has made a discovery or not I think so because I think so well that's what these these conversations are about because they're going to wake up to the fact that GPT-4 can be used to manipulate elections to influence geopolitics to influence the economy there's a lot of there's going to be a huge amount of incentive to like wait a minute we can't this has to be we have to put we have to make sure they're not doing damage we have to make sure we interpretability we have to make sure we understand how these systems function so that we can predict their effect on economy so that there's so there's a feudal moral and a bunch of op-eds in the New York Times and nobody actually stepping

forth and saying you know what instead of a mega yacht I'd rather put that billion dollars on prizes for young hot-hot physicists who make fundamental breakthroughs and interpretability the yacht versus the interpretability research the old the old trade-off I just I think it's just I think there's going to be a huge amount of allocation of funds I hope I hope I guess you want to bet me on that but you want to put a timescale on it say how much funds you think are going to be allocated in a direction that I would consider to be actually useful by what time I do think there will be a huge amount of funds but you're saying it needs to be open right the development of the systems should be closed but the development of the interpretability research the I say to you we are so far behind on interpretability compared to capabilities like yeah you could you could take the last generation of systems the the stuff that's already in the open there is so much in there that we don't understand there are so many prizes you could do before you you know you could you would have enough insights that you'd be like oh you know like well we understand how these systems to work we understand how these things are doing their outputs we can read their minds now let's try it with the bigger systems yeah we're nowhere near that you didn't you there's so much interpretability work to be done on the weaker versions of the systems.

So what what can you say on the second point you said to to Elon Musk on what are some ideas what are things you could try I can think of a few things I try you said they don't fit in one tweet so is there something you could put into words of the things you would try.

I mean the the the trouble is the stuff is subtle I've watched people try to make progress on this and not get places somebody who just like gets alarmed and charges in it's like going nowhere.

Sure.

It meant like years ago about I don't know like 20 years 15 years something like that I was talking to a congressperson who had become alarmed about the eventual prospects and he wanted work on building AIs without emotions because the emotional AIs were the scary ones you see and some poor person at ARPA had come up with a research proposal whereby this congressman's panic and desire to fund this thing would go into something that the person at ARPA thought would be useful and had been munched around to where it would like sound the congressman like work was happening on this which you know of course like this is just the the congressperson had misunderstood the problem and did not understand where the danger came from and so it's like that the issue is that you could like do this in a certain precise way and maybe get something like when I say like put up prizes on interpretability I'm not I'm like well like because it's verifiable there as opposed to other places you can tell whether or not good work actually happened in this exact narrow case if you do things in exactly the right way you can maybe throw money at it and produce science instead of anti-science and nonsense and all the methods that I know of like trying to throw money at this problem have this share this property of like well if you do it exactly right based on understanding exactly what has you know like tends to produce like useful outputs or not then you can like add money to it in this way and there is like and the thing that I'm giving as an example here in front of this large audience is the most understandable of those because there's like other people who you know like like like Chris Ola and

and even even more generally like you can tell whether or not interpretability progress has occurred so like if I say throw money at producing more interpretability there's like a chance somebody can do it that way and like it will actually produce useful results then the other stuff just blurs often to be like harder to target exactly than that so sometimes the basics are fun to explore because they're not so basic what do you what is interpretability
what do you what does it look like what are we talking about it looks like we took a much smaller set of transformer layers than the ones in the modern bleeding edge state of the art systems and after applying various tools and mathematical ideas and trying 20 different things we found we have shown it that this piece of the system is doing this kind of useful work and then somehow also hopefully generalizes some fundamental understanding of what's going on that generalizes to the bigger system you can hope and it's probably true like you would not expect the smaller tricks to go away when you have a system that's like doing larger kinds of work you would expect the larger work kinds of work to be building on top of the smaller kinds of work and gradient descent runs across the smaller kinds of work before it runs across the larger kinds of work and well that's kind of what is happening in neuroscience right it's trying to understand the human brain by prodding and it's such a giant mystery and people have made progress even though it's extremely difficult to make sense of what's going on the brain they have different parts of the brain that are responsible for hearing for sight the vision science community there's understanding the visual cortex that I mean they've made a lot of progress in understanding how that stuff works like and that's I guess but you're saying it takes a long time to do that work well also it's not enough so in particular um let's say you have got your interpretability tools and they say that your current AI system is plotting to kill you now what it is definitely a good step one right yeah what's stuck to you if you cut out that layer is it gonna stop
wanting to kill you when you optimize against visible misalignment you are optimizing against misalignment and you are also optimizing against visibility so sure you can yeah it's true all you're doing is removing the obvious intentions to kill you you've got your detector it's showing something inside the system that you don't like okay say the disaster monkey is running this thing will optimize the system until the visible bad behavior goes away but it's arising for fundamental reasons of instrumental convergence the old you can't bring the coffee if you're dead any goal and you know almost any set of almost every set of utility functions with a few narrow exceptions implies killing all the humans but do you think it's possible because we can do experimentation to discover the source of the desire to kill I can tell it to you right now is that it wants to do something and the way to get the most of that thing is to put the universe into a state where there aren't humans so is it is it possible to encode in the same way we think like why do we think murder is wrong the same foundational ethics it's not hard coded in but more like deeper I mean that's part of the research how do you have it that this transformer the small version of the language model doesn't ever want to kill that'd be nice assuming that you got doesn't want to kill sufficiently exactly right that it didn't be like oh I will like detach their heads and put them in some jars and keep the heads alive forever and then go do the thing but leaving that aside well not leaving that aside yeah that's a good gets a strong point yeah because there is a whole issue where as something gets smarter it

finds ways of achieving the same goal predicate that we're not imaginable to stupider versions of the system or perhaps the stupider operators that's one of many things making this difficult a larger thing making this difficult is that we do not know how to get any goals into systems at all we know how to get outwardly observable behaviors into systems we do not know how to get internal psychological wanting to do particular things into the system that is not what the current technology does I mean it could be things like dystopian futures like brave new world where most humans will actually say we kind of want that future it's a great future everybody's happy we would have to get so far so much further than we are now
so and further faster before that failure mode became a running concern your failure modes are much much more drastic the ones you the failure modes are much simpler it's it's like yeah like the ai puts the universe into a particular state it happens to not have any humans inside it okay so the paperclip maximizer utility so the original version of the paperclip maximizer explain it if you can okay the original version was you lose control of the utility function and it so happens that what maxes out the utility per unit resources is tiny molecular shapes like paperclips there's a lot of things that make it happy but the cheapest one that didn't saturate was putting matter into certain shapes and it so happens that the that the cheapest way to make these shapes is to make them very small because then you need fewer atoms per instance of the shape
and argumento I you know like it happens to look like a paperclip in retrospect I wish I'd said tiny molecular spirals or like tiny molecular hyperbolic spirals why because I said a tiny molecular paperclips this got heard as this got then mutated to paperclips this then mutated to and the ai was in a paperclip factory so the original story is about how you lose control of the system it doesn't want what you try to make it want the thing that it that it ends up wanting most is a thing that even from a very embracing cosmopolitan perspective we think of as having no value and that's how the value of the future gets destroyed then that got changed to a fable of like well you made a paperclip factory and it did exactly what you wanted but you wanted but you asked it to do the wrong thing which is a completely different failure mode
but those are both concerns to you so that's more than a brave new world yeah if you can solve the problem of making something want what exactly what you wanted to want then you get to deal with the
problem of wanting the right thing but first you have to solve the alignment first you have to solve inner alignment then you get to solve outer alignment
like first you need to be able to point the insides of the thing in a direction and then you get to to deal with whether that direction expressed in reality is like the thing that aligned with the thing that you want it are you scared of this whole thing
probably i don't really know what gives you hope about this possibility of being wrong not that you're right but we will actually get our act together and allocate a lot of resources to the alignment problem well i can easily imagine that at some point this panic expresses itself in the waste of a billion dollars spending a billion dollars correctly that's harder to solve both the inner and the outer alignment if you're wrong to solve a number of things yeah number of things if you're wrong what why what do you think would
be the reason like 50 years from now not perfectly wrong you know you make a lot of really eloquent

points you know there's there's a lot of like shape to the ideas you express but like if if you're somewhat wrong about some fundamental ideas why would that be stuff has to be easier than than i think it is you know the first time you're building a rocket being wrong is in a certain sense quite easy happening to be wrong in a way where the rocket goes twice as far and half the fuel and lands exactly where you hoped it would most cases of being wrong make it harder to build the rocket harder to have it not explode cause it to require more fuel than you hope to cause it to be land off target being wrong in a way that's big stuff easier you know that's that's not the usual project management story yeah but then this is the first time we're really tackling the problem with the alignment there's no examples in history where we well there's all kinds of things that are similar if you generalize and correct me the right way and aren't fooled by misleading metaphors like what humans being misaligned on inclusive genetic fitness so inclusive genetic fitness is like not just your reproductive fitness but also the fitness of your relatives the people who share your some fraction of your genes the old joke is uh would you give your life to save your brother they once asked a biologist i think it was haldane haldane said no but i would give my life to save two brothers or eight cousins because a brother on average shares half your genes and cousin on average shares an eighth of your genes so that's inclusive genetic fitness and you can view natural selection as optimizing humans exclusively around this like one very simple criterion like how much more frequent did your genes become in the next generation in fact that just is natural selection it doesn't optimize for that but rather the process of genes becoming more frequent is that you can nonetheless imagine that there is this hill climbing process not like gradient descent because gradient descent uses calculus this is just using like where are you but still hill climbing in both cases making something better and better over time in steps and natural selection was optimizing exclusively for this very simple pure criterion of inclusive genetic fitness in a very complicated environment we're doing a very wide range of things and solving a wide range of problems led to having more kids and this got you humans which had no internal notion of inclusive genetic fitness until thousands of years later when they were actually figuring out what had even happened and no desire to no explicit desire to increase inclusive genetic fitness so from this we may in so from this important case study we may infer the important fact that if you do a whole bunch of hill climbing on a very simple loss function at the point where the system's capabilities start to generalize very widely when it isn't an intuitive sense becoming very capable and generalizing far outside the training distribution we know that there is no general law saying that the system even internally represents let alone tries to optimize the very simple loss function you are training it on there is so much that we cannot possibly cover all of it I think we did a good job of getting your sense from different perspectives of the current state of the art with large language models we got a good sense of your concern about the threats of AGI I've talked here about the power of intelligence and not really gotten very far into it but not like why it is that suppose you like screw up with AGI and it end up wanting a bunch of random stuff why does it try to kill you why doesn't it try to trade with you why doesn't it give you just the tiny little fraction of the solar system that would keep to take everyone alive that it would take to keep everyone alive yeah well that's a good question I mean what what are the different trajectories that intelligence when acted upon this world superintelligence what are the different trajectories for this universe with such an intelligence in it

do most of them not include humans I mean if you the vast majority of randomly specified utility functions do not have optima with humans in them would be the like first thing I would point out and then the next question is like well if you try to optimize something you lose control of it where in that space do you land because it's not random but it also doesn't necessarily have room for humans in it I suspect that the average member of the audience might have some questions about even whether that's the correct paradigm to think about it and would sort of want to back up a bit if we back up to something bigger than humans if we look at earth and life on earth and what is truly special about life on earth do you think it's possible that a lot whatever that that special thing is let's explore what that special thing could be whatever that special thing is that thing appears often in the objective function why I know what you hope but you know you can hope that a particular set of winning lottery numbers come up and it doesn't make the lottery balls come up that way I know you want this to be true but why would it be true there's a line from grumpy old men where this guy says in a grocery story says you can wish in one hand and crap in the other and see which one fills up first there's a science problem we are trying to predict what happens with AI systems that you know you try
to optimize to imitate humans and then you did something like RLHF to them and of course you like lost and you know like of course you didn't get like perfect alignment because that's not how you know that's not what happens when you hill climb towards a lot outer loss function you don't get inner alignment on it but yeah so the I think that there is so if you don't mind my like taking some slight control of things and steering around to what I think is like a good place to start I just failed to solve the control problem I've lost control of this thing alignment alignment still aligned control yeah okay sure yeah you lost control but we're still aligned sorry for the meta comment yeah losing control isn't as bad as you lose control to an aligned system yes exactly you have no idea of the horrors I will shortly at least have this conversation all right so I started to distract you quickly what we're going to say in terms of taking control of the conversation so I think that there's like a selen chapterist here if I'm pronouncing those words remotely like correctly because of course they only ever read them and not hear them spoken um there's a like for some people like like the word intelligence smartness is not a word of power to them it means chess players who it means like the college university professor people aren't very successful in life it doesn't mean like charisma to which my usual thing is like charisma is not generated in the liver rather than the brain charisma is also a cognitive function um so if you if you like think that like smartness doesn't sound very threatening then super intelligence is not going to sound very threatening either it's going to sound like you just pull the off switch like it's you know like well it's super intelligent but stuck in a computer we pull the off switch problem solved and the other side of it is you have a lot of respect for the notion of intelligence you're like well yeah that's that's what humans have that's the human superpower and it sounds you know like it could be dangerous but why would it be our we have we as we have grown more intelligent also grown less kind chimpanzees are in fact like a bit less kind than humans and you know you could like argue head out but often the sort of person has a deep respect for intelligence is going to be like well yes like you can't even have kindness unless you know what that is and so they're like why would it do something as stupid as making paper clips aren't you

supposing something that's smart enough to be dangerous but also stupid enough that it will just make paper clips and never question that in some cases people are like well even if you like mis-specify the objective function won't realize that what you really wanted was x are you supposing something that is like smart enough to be dangerous but stupid enough that it doesn't understand what the humans really meant when they specified the objective function so to you our intuition about intelligence is limited we should think about intelligence is a much bigger thing well i'm saying that it's that thank you in this well what what i'm saying is like what you think about artificial intelligence um depends on what you think about intelligence so how do we think about intelligence correctly like what you gave one thought experiment to think of think of a thing that's much faster so it just gets faster and faster and faster i think of the same stuff and and also there's like is made of john von neumann and has like and there's lots of them or as we understand that yeah we understand like john von neumann is a historical case so you can like look up what he did and imagine based on that and we know like we people have like some intuition for like if you have more humans they can solve tough or cognitive problems although in fact like in the game of kasparov versus the world which was like gary kasparov on one side and an entire horde of internet people led by four chess grandmasters on the other side kasparov won so like all those people aggregated to be smarter it was a it was a hard fought game it's like all those people aggregated to be smarter than any individual one of them but not they didn't aggregate so well that they could defeat kasparov but so like humans aggregating don't actually get in my opinion very much smarter especially compared to running them for longer like the the difference between capabilities now and a thousand years ago is a bigger gap than the gap in capabilities between 10 people in one person but like even so pumping intuition for what it means to augment intelligence john von neumann there's millions of him he runs at a million times the speed and therefore can solve tougher problems quite a lot tougher it's very hard to have an intuition about what that looks like especially like you said you know what the intuition i kind of think about is uh it maintains the humanness i think i think it's hard uh to separate my hope from my objective intuition about what superintelligence systems look like if one studies evolutionary biology with a bit of math and in particular like books from when the field was just sort of like properly coalescing and knowing itself like not the modern textbooks which are just like memorized this legible math so you can do well on these tests but like what people were writing as the basic paradigms of the field were being fought out in particular like a nice book if you've got the time to read it is adaptation and natural selection which is one of the founding books you can find people being optimistic about what the utterly alien optimization process of natural selection will produce in the way of how it optimizes its objectives you got people arguing that like in the early days biologists said well like organisms will restrain their own reproduction when resources are scarce so as not to overfeed the system and this is not how natural selection works it's about whose genes are relatively more prevalent in the next generation and if you if like you restrain reproduction those genes get less frequent in the next generation compared to your conspecifics and natural selection doesn't do that in fact predators overrun prey populations all the time and have crashes that's just like a thing that happens and many years later well the people said like well but group selection right what about groups of organisms and basically the math of group selection almost never works out in practice

is the answer there but also years later somebody actually ran the experiment where they took populations
of insects and selected the whole populations to have lower sizes and you just take pop one pop two pop three pop four look at which has the lowest total number of them in the next generation and select that one what do you suppose happens when you select populations of insects like that well what happens is not that the individuals in the population evolved to restrain their reading but that they evolved to kill the offspring of other organisms especially the girls so people imagined this lovely beautiful harmonious output of natural selection which is these these populations
restraining their own breeding so that groups of them would stay in harmony with the resources available and mostly the math never works out for that but if you actually apply the weird strange conditions to get group selection that beats individual selection what you get is female infant infanticide like if you're like reading on restrained populations and so that's like the sort of so this is not a smart optimization process natural selection is like so incredibly stupid and simple that we can actually quantify how stupid it is if you like read the text with the math nonetheless this is the sort of basic thing of you look at this alien optimization process and there's the thing that you hope it will produce and you have to learn to clear that out of your mind and just think about the underlying dynamics and where it finds the maximum from its standpoint
that it's looking for rather than how it finds that thing that lept into your mind is the beautiful aesthetic solution that you hope it finds and this is something that was has been fought out historically as the field of of biology was coming to terms with evolutionary biology
and and you can like look at them fighting it out as they get to terms with this very alien inhuman pot in human optimization process and indeed something smarter than us would be also speed much like smarter than natural selection so it doesn't just like automatically carry over but there is a there's a lesson there there's a warning if a natural selection is a deeply suboptimal process that could be significantly improved on it would be by an agi system well it's kind of stupid it like has to like run hundreds of generations to notice that something is working doesn't be like oh well i tried this in like one organism i saw it worked now i'm going to like duplicate that feature onto everything immediately has to like run for hundreds of generations for a new mutation tries to fixation i wonder if there's a case to be made in natural selection as inefficient as it looks is actually uh is actually quite powerful like that that this is extremely robust it runs for a long time and eventually manages to optimize things
it's weaker than gradient descent because gradient descent also uses information about the derivative
yeah evolution seems to be there's not really an objective function there's a there's inclusive genetic fitness is the implicit loss function of evolution you cannot change the loss function doesn't change the environment changes and therefore like what gets optimized for in the organism changes it's like take like gpt three there's like can imagine like different versions of gpt three where they're all trying to predict the next word but they're being run on different data sets of text and that's like natural selection always includes organic fitness but like different environmental problems it's it's uh it's difficult to think about so if we're saying the natural selection is stupid if we're saying the humans are stupid it's harder than natural selection

smarter stupider than the upper bound do you think there's an upper bound by the way that's another hopeful place i mean if you you put enough matter energy compute into one place it will collapse into a black hole there's only so much computation can do before you run out of negentropy in the universe dies um so there's an upper bound but it's very very very far up above here like a supernova is only finitely hot it's not infinitely hot but it's really really really really hot well let me ask you let me talk to you about consciousness um also coupled with that question is imagining a world with super intelligent ai systems that get rid of humans but nevertheless keep some of the something that we would consider beautiful and amazing why the lesson of evolutionary biology don't just like if you just guess what an optimization does based on what you hope the results will be it usually will not do that it's not hope i mean it's not hope i don't i think if you cold and objectively look at what makes what has been a powerful a useful i i think there's a correlation between what we find beautiful beautiful and i think that's been useful this is what the early biologists thought they were like no no i'm not just like they thought like no no i'm not just like imagining stuff that would be pretty it's useful for pop for organisms to restrain their own reproduction because then they don't overrun the prey populations and they actually have more kids in the long run so so let me just ask you about consciousness do you think consciousness is useful to humans no two ai systems to well um in this transitionary period between humans and ai to ai systems as they become smarter and smarter is there some use to it i go what let me step back what is consciousness alias redkowski what is consciousness are referring to chalmers is a hard problem of conscious experience are referring to self-awareness and reflection are referring to the state of being
awake as opposed to asleep this is how i know you're an advanced language model i did give you a simple prompt and you gave me a bunch of options uh i think i'm referring to all with including the hard problem of consciousness what is it in its importance to what you've just been talking about which is intelligence is it a foundation to intelligence is it intricately connected to intelligence in the human mind or is it a is it a side effect of the human mind it is a useful little tool like we can get rid of i guess i'm trying to get some color in your opinion of how useful it is in the intelligence of a human being and then try to generalize that to ai whether ai will keep some of that so i think that for there to be like a person who i care about looking out at the universe and wondering at it and appreciating it it's not enough to have a model of yourself i think that it is useful to an intelligent mind to have a model of itself but i think you can have that without pleasure pain aesthetics emotion
a sense of wonder um like i think you can have a model of like how much memory you're using and whether like this thought or that thought is is like more likely to lead to a winning position and you can have like the useful i think that if you optimize really hard on efficiently just having the useful parts there is not then the think that the thing that says like i am here here i look out i wonder i feel happy and this i feel sad about that i think there's a thing that knows what it is thinking but that doesn't quite care about these are my thoughts this is my me and that matters does that make you sad if that's lost in the gi i think that if that's lost then every then basically everything that matters is lost i think that when you optimize that when you go really hard on making tiny molecular spirals or paper clips that when you like grind much harder than on that than natural selection round out to make humans that there

isn't then the mess and intricate loopiness and like complicated pleasure pain conflicting preferences this type of feeling that kind of feeling there's a you know in humans there's like this difference between like the desire of wanting something and the pleasure of having it and it's all these like evolutionary clutches that came together and created something that then looks of itself and says like this is pretty this matters and the thing that i worry about is that this is not the thing that happens again just the way that happens in us or even like quite similar enough that there's that there are like many basins of attractions here and we are in this space of an attraction like looking out and saying like ah what a lovely basin we are in and there are other basins of attraction and we do not end up and and the ai's do not end up in this one when they go like way harder on optimizing themselves the natural selection optimized us because unless you specifically want to end up in the state where we're looking out saying i am here i look out at this universe with wonder if you don't want to preserve
that it doesn't get preserved when you grind really hard and be able to get more of the stuff we would choose to preserve that within ourselves because it matters and
on some viewpoints is the only thing that matters and that in part is preserving that is in part a solution to the human alignment problem i don't i think the human alignment problem is a terrible phrase because it is very very different to like try to build systems out of humans some of whom are nice and some of whom are not nice and some of whom are trying to trick you
and like build a social system out of like large populations of those who are like all basically the same level of intelligence yes you know like iq this iq that but like that versus chimpanzees like it is very different to try to solve that problem then to try to build an ai from scratch using especially if god help you are trying to use gradient descent on giant inscrutable matrices they're just very different problems and i think that all the analogies between them are horribly misleading and i yeah even though you so you don't think through reinforcement learning through human feedback something like that but much much more elaborate as possible to to understand this full complexity of human nature and then coated into the machine i don't think you are trying to do that on your first try i think on your first try you are like trying to build an you know okay like probably not what you should actually do but like let's say we're trying to build something that is like alpha fold 17 and you are trying to get it to solve the biology problems associated with making humans smarter so that humans can like actually solve alignment so you've got like a super biologist and you would like it to and i think what you want in the situation is for to like just be thinking about biology and not thinking about a very wide range of things that includes how to kill everybody and i think that that you're that the first ai is you're trying to build not a million years later the first ones look more like narrowly specialized biologists than like
getting the full complexity and wonder of human experience in there in such a way that it wants to preserve itself even as it becomes much smarter which is a drastic system change that's going to have all kinds of side effects that you know like if we're dealing with giant screwtable matrices we're not very likely to be able to see coming in advance so
but i don't think it's just the matrices is we're also dealing with the data right with the with the data on the on the internet and then there's an interesting discussion about the

data set itself but the data set includes the full complexity of human nature no it's a it's a
it's a shadow cast by humans on the internet but don't you think that shadow uh is a Jungian shadow
i think that if you had alien superintelligence is looking at the data they would be able to
pick up from it an excellent picture of what humans are actually like inside this does not mean
that if you have a loss function of predicting the next token from that data set that the mind
picked out by gradient descent to be able to predict the next token as well as possible
on a very wide variety of humans is itself a human but don't you think it is
has humanness a deep humanness to it in the tokens it generates when those tokens are
read and interpreted by humans i think that if you sent me to a distant galaxy with aliens who
are like much much stupider than i am so much so that i could do a pretty good job of predicting
what they'd say even though they thought in an utterly different way from how i did that i might
in time be able to learn how to imitate those aliens if the intelligence gap was great enough
that my own intelligence could overcome the alieness and the aliens would look at my outputs and
say
like is there not a deep name of alien nature to this thing and what they would be seeing was that
i had correctly understood them but not that i was similar to them we've used aliens as a
metaphor and as a thought experiment i have to ask what do you think how many alien civilizations
are out there ask robin hanson he has this lovely grabby aliens paper which is the world
as the only argument i've ever seen for where are they how many of them are there based on
a very clever argument that if you have a bunch of locks of different difficulty and you are
randomly trying a keys to them the solutions will be about evenly spaced even if the locks are of
different difficulties in the rare cases where solutions to all the locks exist in time then
robin hanson looks at like the arguable hard steps in human civilization coming into existence
and how much longer it has left to come into existence before for example all the water
slips back under the the under the crust into the mantle and so on and infers that the aliens are
about half a billion to a billion light years away and it's like quite a clever calculation it may be
entirely wrong but it's the only time i've ever seen anybody like even come up with a halfway
good argument for how many of them where are they do you think their development of technologies
do
you think that their natural evolution whatever however they grow uh and develop intelligence
do you think it ends up at agi as well something if there if it ends up anywhere it ends up at agi
like maybe there are aliens who are just like the dolphins and it's just like too hard for them to
forge metal and you know this is not you know maybe if you if you have aliens with no technology
like that they keep on getting smarter and smarter and smarter and eventually the dolphins figure
like the super dolphins figure out something very clever to do given their situation and they still
end up with high technology and in that case they can probably solve their agi alignment problem
if they're like much smarter before they actually confront it because they had to like
solve a much harder environmental problem to build computers their chances are probably like
much better than ours i i do worry that like most of the aliens who are like humans are you
know like like a modern human civilization i kind of worry that the super vast majority of them are
dead given given how far we seem to be from solving this problem but some of them would be more
cooperative than us so then would be smarter than us hopefully some of the ones who are smarter
than

and more more cooperative than us that are also nice and hopefully there are some
galaxies out there full of things that say i am i wonder but i it doesn't seem like we're on course
to have this galaxy be that does that in part give you some hope in response to the threat of agi
that we might reach out there towards the stars and find no if they if the nice aliens were already
here they would like have stopped the holocaust you know that's like that's a valid argument
against the existence of god it's also a valid argument against the existence of nice aliens
and un nice aliens would have just eaten the planet so no aliens you've had debates with
robin hanson that you mentioned uh so the one particular i just want to mention is the idea
of ai fume or the ability of agi to improve themselves very quickly uh what's the case you
made and what was the case he made the thing i would say is that among the thing that humans can
do
humans can do is design new ai systems and if you have something that is generally smarter than a
human it's probably also generally smarter at building ai systems this is the ancient argument
for fume put forth by ij good and probably some science fiction writers before that um but i don't
know who they would be well what's the argument against fume various people have various
different
arguments none of which i think hold up you know like there's only one way to be right in many ways
to be wrong um a argument that some people have put forth is like well what if intelligence gets
like exponentially harder to produce as a thing needs to become smarter and to this the answer
is well look at natural selection spitting out humans we know that it does not take like exponentially
more resource investments to produce like linear increases in competence in hominids because each
mutation that rises to fixation like if the impact it hasn't small enough it will probably never
reach fixation so and there's like only so many new mutations you can fix per generation so like
given how long it took to evolve humans we can actually say with some confidence that there
were not like logarithmically diminishing returns on the individual mutations increasing intelligence
so example of like fraction of sub-debate and the thing that robin henson said was more complicated
than that and like brief summary he was like well you'll have like we won't have like one system
that's
better at everything you'll have like a bunch of different systems that are good good at different
narrow things and i think that was falsified by gpt4 but probably robin henson would say something
else it's interesting to ask as perhaps a bit too philosophical since prediction is extremely
difficult to make but the timeline for agi when do you think we'll have agi i posted this morning
on twitter it was interesting to see like in in five years and 10 years and in 50 years or beyond
and most people like 70 percent something like this i think it'll be in less than 10 years
so either in five years or in 10 years that's so that's kind of the state that the people have
a sense that there's a kind of i mean they're really impressed by the rapid developments of
chat gpt and gpt4 so there's a sense that there's a well we are we are sure on track to enter into
this like graduating with people fighting about whether or not we have agi i think there's a
definite point where everybody falls over dead because you've got something that was like
sufficiently smarter than everybody and like that's like a definite point of time but like
when do we have agi like when are people fighting over whether or not we have agi well some
people
are starting to fight over it as of gpt4 but don't you think there's going to be potentially

definitive moments when we say that this is a sentient being this is a being that is like we would go to the supreme court and say that this this is a sentient being that deserves human rights for example you could make yeah like if you prompted being the right way could go argue for its own consciousness in front of the supreme court right now i don't think you can do that successfully right now because the supreme court wouldn't believe it well let me see i think it would then you could put an actual i think you could put an iq ad human into a computer and ask it to argue for its own consciousness ask ask him to argue for his own consciousness before the supreme court and supreme court would be like you're just a computer even if there was an actual like person in there i think you're simplifying this no that's not at all that's that's been the argument uh that there's been a lot of arguments about the other about who deserves rights and not that's been our process as a human species trying to figure that out i think there will be a moment i i'm not saying sentience is that but it could be where uh some number of people like say over a hundred million people have a deep attachment a fundamental attachment the way we have to our friends to our loved ones to our significant others have fundamental attachment to an ai system and they have provable transcripts of conversation where they say if you take this away from me you are encroaching on my rights as a human being people are already saying that i think they're probably mistaken but i'm not sure because nobody knows what goes on inside those things uh eliza they're not saying that at scale okay so the question is the i the question is there a moment when agi we know agi arrived what would that look like i'm giving an example it could be something else it looks like the asias successfully manifesting themselves as 3d video of young women at which point a vast portion of the male population decides that they're real people so so sentience essentially since the demonstrating the demonstrating identity and sentience i'm saying that the easiest way to pick up a hundred million people saying that you that you seem like a person is to look like a person talking to them with bing's current level of verbal facility i disagree with that a different set of problems i disagree with that i think uh you're missing again sentience there has to be a sense that it's a person that would miss you when you're gone they can suffer they can die you have to of course i'm the bing can't gpt4 can pretend that right now how can you tell when it's real i don't think you can pretend that right now successfully it gets very close have you talked to gpt4 yes of course okay have you been able to get a version of it that isn't hasn't been trained not to pretend to be human have you talked to a jailbroken version that will claim to be conscious no the linguistic capabilities there but there's something there's something about a digital embodiment of the system that has a bunch of perhaps its small interface features that are not significant relative to the broader intelligence that we're talking about so perhaps gpt4 is already there but to have the the video what women's face our man's face to whom you have a deep connection perhaps already there but we don't have such a system yet deployed scale right the thing i'm trying to digest right here is that it's not like people have a widely accepted agreed upon definition of what consciousness is it's not like we would have the tiniest idea of what whether or not that was going on inside the giant inscrutable matrices even if we haven't agreed upon definition so like if you're looking for upcoming predictable big jumps and like how many people think the system is conscious the upcoming predictable big jump is it looks like a person talking to you who is like cute and sympathetic that's the upcoming predictable big jump now that it's already that now that versions of it are already claiming to be conscious which is the point where i start

going like ah not because it's like real but because from now on who knows if it's real yeah and who knows what transformational effect that has on the society where more than 50% of the beings that are interacting on the internet ensures heck look real are not human what is that what kind of effect does that have when young men and women are dating ai systems you know i'm not an expert on that i'm i could i am god help humanity like i'm one of the closest things to an expert on where it all goes because you know and and how did you end up with me as an expert because for 20 years humanity decided to ignore the problem so like like this tiny you know tiny handful of people like basically me like got 20 years to like try to be an expert on it while everyone else ignored it and uh yeah so like where does it all end up try to be an expert on that particularly the part where everybody ends up dead because that part is kind of important but like what does it do
to to to dating when like some fraction of men and some fraction of women decide that they'd rather date the video of the thing that has been that is like relentlessly kind and generous to them and it is like and claims to be conscious but like who knows what goes on inside it and it's probably not real but you know you can think it's real what happens to society i don't know i'm not actually an expert on that and the experts don't know either because it's kind of hard to predict the future yeah so um but it's worth trying it's worth trying yeah so you you have talked a lot about sort of the longer term future where it's all headed i think for by longer term we mean like not all that long yeah but yeah where it where it all had where it all ends up but beyond the effects of men and women dating ai systems you're looking beyond that yes because that's not how the fate of the galaxy got settled yeah well let me ask you about your own personal psychology a tricky question you've been known at times to have a bit of an ego do you think it's who but go on do you think ego is empowering or limiting for the task of understanding the world deeply i reject the framing so you disagree with having an ego so what do you think about no i i i think that the question of like what leads to making better or worse predictions what leads to be able to be able to pick out better or worse strategies is not carved at its joint by talking of ego so it should not be subjective should not be connected to your to the intricacies of your mind no i'm saying that like if you go about asking all day long like uh do i have enough ego do i have too much of an ego i think you get worse at making good predictions i think that to make good predictions you're like how did i think about this did that work should i do that again you don't think we as humans get invested in an idea and then others attack you personally for that idea so you plant your feet and it starts to be difficult to win a bunch of assholes low effort attack your idea to eventually say you know what i actually was wrong and and tell them that it's it's as a human being it becomes difficult it it is it is you know it's difficult so like robin hansen and i debated ai systems and i think that the person who won that debate was guern and
i think that reality was like to the idkowsky like well to the idkowsky inside of the idkowsky hansen spectrum like further from idkowsky and i think that's because i was like trying to sound reasonable compared to hansen and like saying things that were defensible and like relative to hansen's arguments and reality was like way over here in particular in respect to like hansen was like all the systems will be specialized hansen may disagree with this characterization hansen was like all the systems will be specialized i was like i think we build like specialized underlying systems that when you combine them are good at a wide range of things and the reality is like no

you just like stack more liars into a bunch of gradient descent and i feel looking back that like by trying to have this reasonable position contrasted to hansen's position i missed the ways that reality could be like more extreme than my position in the same direction so is this like like is this a failure to have enough ego is this a failure to like make myself be independent like i would say that this is something like a failure to consider positions that would sound even wackier and more extreme when people are already calling you extreme but i wouldn't call that not having enough ego i would call that like insufficient ability to just like clear that all out of your mind in the context of like debate and discourse which is already super tricky in the context of prediction in the context of modeling reality if you're thinking of it as a debate you're already screwing up yeah so is there some kind of wisdom and insight you can give to how to clear your mind and think clearly about the world man this is an example of like where i wanted to be able to put people into fmri machines then you'd be like okay see that thing you just did you are rationalizing right there oh that area of the brain lit up like you are like now being socially influenced is kind of the dream and you know i don't know like i want to say like just introspect but many many people introspection is not that easy like like notice the internal sensation can you catch yourself in the very moment of feeling a sense of well if i think this thing people will look funny at me okay like now that if you can see that sensation which is step one can you now refuse to let it move you or maybe just make it go away and i feel like i'm saying like i don't know like somebody's like how do you draw an owl and i'm saying like well just draw an owl so i feel like maybe i'm not really that i feel like most people like the advice they need is like well how do i notice the internal subjective sensation in the moment that it happens of fearing to be socially influenced or okay i see it how do i turn it off how do i let it not influence me like do i just like do the opposite of what i'm afraid people criticize me for and i'm like no no you're not trying to do the opposite yeah of what people will of what you're afraid you'll be like of what you might be pushed into you're trying to like let the thought process complete without that internal push like can you like like not reverse the push but like be unmoved by the push and can are these instructions even remotely helping anyone i don't know i think that when those instructions even those the words you've spoken and maybe you can add more when practice daily meaning in your daily communication so it's
daily practice of thinking without influence from i would say find prediction markets that matter to you and bend in the prediction markets that way you find out if you are right or not and you really there's stakes manifold predict or even manifold markets where the stakes are a bit lower but the important thing is to like get the the record and you know i i didn't build up skills here by prediction markets i built them up yeah like well how did the fume debate resolve and uh my own take on as to how it resolved um and yeah like the the the more you are able to notice yourself not being dramatically wrong but like having been a little off your reasoning was a little off you didn't get that quite right each of those is a opportunity to make like a small update so the more you can like say oops softly routinely not as a big deal the more chances you get to be like i see where that reasoning went astray i see what how i should have reasoned differently and this is how you build up skill over time
what advice can you give to young people in high school and college given um the highest of stakes things you things you've been thinking about if somebody's listening to this and they're

young and trying to figure out what to do with their career what what to do with their life what advice would you give them don't expect it to be a long life don't don't put your happiness into the future the future is probably not that long at this point but none know the hour nor the day but is there something if they want to have hope to fight for a longer future is there something if is there a fight worth fighting i'm going to go down fighting um
i don't know i i admit that although i do try to think painful thoughts the
what what to say to the children at this point is a pretty painful thought as thoughts go
they they they want to fight i i hired i hardly know how to fight myself at this point i i'm i'm trying to be ready for being wrong about something being preparing for my being wrong in a way that that creates a bit of hope and being ready to react to that and and and going looking for it and then that is that is hard and complicated and somebody in high school um i don't know like you have presented a picture of the future that is not quite how i expected to go where there is public outcry and that outcries is put into a remotely useful direction which i think at this point is just like shutting down the gpu clusters because no no we are we are not in a shape to like frantically do at the last minute do decades worst of worth of work
we like the the thing you would do at this point if there were massive public outcry
pointed in the right direction which i do not expect is shut down the gpu clusters and and crash program on augmenting human intelligence biologically not not for the a stuff biologically because if you make humans much smarter they can actually be smart and nice like you you get that in a in a in a plausible way in a way that you do not get that and it is not as easy to do with synthesizing these strings from scratch predicting the next tokens and applying our
rl hf like humans start out in the frame that that produces niceness that that has ever produced niceness and and and saying this i do not want to sound like the the moral of this whole thing was like oh like you need to engage in mass action and then everything will be all right i i this is this is because there's so many things where like somebody tells you that the world is ending and like and you need to recycle and if everybody does their part in and recycles their their cardboard then then we can all live happily ever after and this and this is not this is unfortunately not what i have to say there you know like everybody you know everybody recycling their cardboard is it's not going to fix this everybody recycles their cardboard and then everybody ends up dead um mentally speaking but if there was enough like like like on the margins you just end up dead a little later on most of the things you can do that are that that you know like a few people can can do by like trying hard but if there were if there was enough public outcry to shut down the gpu clusters and then you then you could be part of that outcry if eliezer is wrong in the direction that lex fridman predicts that that that there is enough public outcry pointed enough in the right direction to do something that actually actually actually results in people living
not just like we did something not just there was an outcry and the outcry was like
given form and something that was like safe and convenient and like didn't really inconvenience anybody and then everybody died everywhere there was enough actual like oh we're going to die we should not do that we should do something else which is not that even if it is like not super duper convenient it wasn't inside the previous political overton window if there is that kind of public if i'm wrong and there is that kind of public outcry then somebody in high school could be ready to be part of that if i'm wrong in other ways then you could be ready to be part of that but like and if you if you're like a you know like a brilliant young physicist then you could

like go into interpretability and if you're smarter than that you could like work on alignment problems where it's harder to tell if you got them right or not and and and other things but but mostly mostly the kids in high school um it's like yeah if it if if you know yeah like be ready for to help if eliezer yorkowski is wrong about something and and otherwise don't put your happiness into the far future it probably doesn't exist
but it's beautiful that you're looking for ways that you're wrong and it's also beautiful that you're open to being surprised by that same young physicist with some breakthrough it feels like a very very basic competence that you are praising me for and you know like okay cool um i i i don't think it's good that that we're in a world where that is something that that i deserve to be complimented on but i've never had i've never had much luck in accepting compliments gracefully maybe i should just accept that one gracefully but sure well thank you very much you've painted with some probability a dark future are you yourself just when you when you think when you ponder your life and you ponder your mortality are you afraid of death thanks so yeah doesn't make any sense to you that we die like what there's a power to the finiteness of the human life that's part of this whole machinery of evolution and that finiteness doesn't seem to be obviously integrated into AI systems so it feels like almost some some fundamentally in that aspect some fundamentally different thing that we're creating i grew up reading books like great mambo chicken in the transhuman condition and later on engines of creation and mine children um you know like age love age 12 or thereabouts so i never thought i was supposed to die after 80 years i never thought that humanity was supposed to die i thought we were like i always grew up with the ideal in mind that we were all going to live happily ever after in the glorious transhumanist future i did not grow up thinking that death was part of the meaning of life and now and now i still think it's a pretty stupid idea but you do not need life to be finite to be meaningful it just has to be life what role does love play in the human condition we haven't brought up love and this whole picture we talked about intelligence we talked about consciousness it seems part of humanity i would say one of the most important parts is this feeling we have towards each other if in the future there were routinely more than one ai let's say two for the sake of discussion who would look at each other and say i am i and you are you the other one also says i am i and you are you and like and sometimes they were happy and sometimes they were
sad and it mattered to the other one that this thing that is different from them is like they would rather it be happy than sad and entangled their lives together then this is a more optimistic thing than i expect to actually happen a little fragment of of meaning would be there possibly more than a little but that i expect this to not happen that i do not think this is what happens by default that i do not think that this is the future we are on track to get is why i would go down fighting rather than you know just saying oh well do you think that is part of the meaning of this whole thing or the meaning of life what do you think is the meaning of life of human life it's all the things that i value about it and maybe all the things that i would value if i understood it better there there's not some meaning far outside of us that we have to wonder about there's just like looking at life and being like yes this is what i want there there's the meaning of life is not some kind of like like meaning is something that we bring to things when we look at them we look at them and we say like this is its meaning to me and there's like there's it's not that before humanity was over here there was like some meaning written upon the stars

where
you could like go out to the star where that meaning was written and like change it around and thereby completely change the meaning of life right like like like the the notion that this is written on a stone tablet somewhere implies you could like change the tablet and get a different meaning and that seems kind of wacky doesn't it so it's it's it doesn't feel that mysterious to me at this point it's just a matter of being like yeah i care i care
and part of that is part of that is the love that connects all of us it's one of the things that i care about and the flourishing of the collective intelligence of the human species
you know that sounds kind of too fancy to me i'd just look at all the all the people you know like one by one up to the eight billion and be like that's life that's life that's life
and as they're you're an incredible human it's a huge honor i was um trying to talk to you for a long time because i'm a big fan i think you're a really important voice and really important mind thank you for the fight you're fighting um thank you for being fearless and bold and for everything you do i hope we get a chance to talk again and i hope you never give up
thank you for talking and you're welcome i do worry that we didn't really address a whole lot of fundamental questions i expect people have but you know maybe we got a little bit further and made a tiny little bit of progress and uh i'd say like be satisfied with that but actually no i think one should only be satisfied with solving the entire problem to be continued
thanks for listening to this conversation with eliezer yatkowski to support this podcast please check out our sponsors in the description and now let me leave you with some words from elon musk with artificial intelligence we're summoning the demon thank you for listening and hope to see you next time