The following is a conversation with Max Tegmark, his third time in the podcast.
In fact, his first appearance was episode number one of this very podcast.
He is a physicist and artificial intelligence researcher at MIT, co-founder of FutureLeft Institute and author of Life 3.0, being human in the age of artificial intelligence.
Most recently, he's a key figure in spearheading the open letter calling for a six-month pause on giant AI experiments like training GPT-4. The letter reads,
we're calling for a pause on training of models larger than GPT-4 for six months.
This does not imply a pause or ban on all AI research and development,
or the use of systems that have already been placed on the market.
Our call is specific and addresses a very small pool of actors who possess this capability.
The letter has been signed by over 50,000 individuals, including 1800 CEOs and over 1500 professors. Signatories include Joshua Benjo, Stuart Russell, Elon Musk, Steve Wozniak, you all know a Harari, Andrew Yang, and many others. This is a defining moment in the history of human civilization, where the balance of power between human and AI begins to shift.
And Max's mind and his voice is one of the most valuable and powerful in a time like this.
His support, his wisdom, his friendship has been a gift I'm forever deeply grateful for.
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 Notion for Project and Team Collaboration, Inside Tracker for Biological Data, and Indeed for Hiring. Choose wisely, my friends. Also, speaking of hiring, if you want to work with our amazing team, or always hiring, whether it's through Indeed or otherwise, go to lexfreedman.com slash hiring. And now onto the full ad reads.
As always, no ads in the middle. I try to make this interesting, but if you must skip them, please still check out our sponsors. I enjoy their stuff. Maybe you will too. This show is brought to you by Notion. I've spoken endlessly about how amazing Notion is, how everybody, all the cool kids are recommending it for just basic note taking. But there's so, so much more.
It's the collaborative aspect of it, the project management aspect of it, the wikis, the document sharing, all of that, all in a simple, powerful, beautifully designed solution.
What can I say? On top of this, there's the Notion AI tool. This is the best integration of large language models into a productivity note taking tool. There are so many amazing features. I mean, it's just endless. Go to the website. You can generate entire presentations and reports based on a to-do list. You can summarize stuff, you can short stuff, you can generate tables based on the description. You can write a summary, you can expand the text, you can change the style of the text, you can fix spelling and grammar, you can translate, you can use simpler language, more complicated language, change the tone of the voice, make it shorter, longer, like I said, everything. It's just so easy to play around with and all of it, no matter what you're doing, will challenge you to think how you write. It will challenge you to expand the style of your writing. It will save you a lot of time, of course, but I just think it makes you a better thinker and productive being in this world. I think that's such a great integration of AI into the productivity workflow. To me, it's not enough for a large language model to be effective at answering questions and having good dialogue. You have to really integrate it into the workflow and Notion, better than anybody else I've seen, has done that. So if that's interesting to you, Notion AI helps you work faster, write better, and think bigger doing tasks that normally take you hours and just minutes. Try Notion AI for free when you go to Notion.com slash Lex. That's

all lowercase Notion.com slash Lex to try the power of Notion AI today. This show is also brought to you by Inside Tracker, a service I use to track biological data. It's really good to do that kind of thing regularly to look at all the different markers in your body to understand what could be made better through lifestyle and through diet changes. It's kind of obvious that decisions about your life should be made based on the data that comes from your body. Not some kind of population study, although those are good. Not some spiritual guru, although those are good. Not some novel, whether it's Harry Potter or Dostoevsky, although those are sometimes good. Not your relative who says, I heard a guy say that a guy does this thing that is very brosalicy sounding. Although sometimes it turns out to be pretty effective. Overall, the best decisions about your life should be based on the things that come from your own body. Inside Tracker uses algorithms to analyze
your blood data, DNA data, data, fitness tracker, all that kind of stuff to give you recommendations. You should be doing it. You should be doing it regularly. So it's not just a one time thing, but regularly over time you see what changes led to improvements in the various markers that come from your body. Get special savings for a limited time when you go to insidetracker.com. This show is also brought to you by Indeed, a hiring website. I think the most important thing in life, not to quote Conan the Barbarian because that would be very inappropriate to quote at this moment. And it's not actually accurate at all. As a reflection, what's important in life, it's only has comedic value. What I really want to say about what's important in life is the people you surround yourself with. And we spend so much of our time in the workplace seeking solutions to very difficult problems together, passionately pursuing ambitious goals, sometimes impossible goals. That is the source of meaning, a sort of a happiness for people. And I think part of that happiness comes from the collaboration with other human beings, the sort of professional depth of connection that you have with other human beings of being together through the grind and surviving and accomplishing the goal or failing in a big epic way, knowing that you have tried together. And so doing that with the right team, I think is one of the most important things in life. So you should surround yourself with the right team. If you're looking to join a team, you should be very selective about that. Or if you're looking to hire a team, you should be very selective about that and use the best tools of the job. I've used Indeed many, many times throughout my life for the teams I've led. Don't overspend on hiring. Visit Indeed.com slash Lex to start hiring now. That's the Indeed.com slash Lex terms and conditions apply. This is the Lex Friedman podcast. To support it, please check out our sponsors in the description. And now, dear friends, here's Max Tagmark.

You were the first ever guest on this podcast, episode number one. So first of all, Max, I just have to say thank you for giving me a chance. Thank you for starting this journey. It's been an incredible journey. Just thank you for sitting down with me and just acting like I'm somebody who matters that I'm somebody who's interesting to talk to. And thank you for doing it. I meant a lot. Thanks to you for putting your heart and soul into this. I know when you delve into controversial topics, it's inevitable to get hit by what Hamlet talks about the slings and arrows and stuff. And I really admire this. It's in an era, you know, where YouTube videos are too long and now it has to be like a 20 minute TikTok, 20 second TikTok clip. It's just so refreshing to see you going exactly against all of the advice and doing these really long form things. And the people appreciate it. You know, reality is nuanced. And thanks for sharing it that way.

So let me ask you again, the first question I've ever asked on this podcast, episode number one, talking to you, do you think there's intelligent life out there in the universe? Let's revisit that question. Do you have any updates? What's your view when you look out to the stars? So when we look at the stars, if you define our universe the way most astrophysicists do, not as all of space, but the spherical region of space that we can see with our telescopes from which light has a time to reach us since our Big Bang. I'm in the minority. I estimate that we are the only life in this spherical volume that has invented internet radios gotten our level of tech.
And if that's true, then it puts a lot of responsibility on us to not
mess this one up. Because if it's true, it means that life is quite rare. And we are stewards of this one spark of advanced consciousness, which if we nurture it and help it grow, it eventually life can spread from here out into much of our universe. And we can have this just amazing future. Whereas if we instead are reckless with the technology we build and just snuff it out due to the stupidity or infighting, then maybe the rest of cosmic history in our universe was just going to be a play for empty benches. But I do think that we are actually very likely to get visited by aliens, alien intelligence quite soon. But I think we are going to be building that alien intelligence. So we're going to give birth to an intelligent alien civilization. Unlike anything that human, the evolution here on Earth was able to create in terms of the path, the biological
path it took. Yeah, and it's going to be much more alien than a cat or even the most exotic animal on the planet right now. Because it will not have been created through the usual Darwinian competition where it necessarily cares about self-preservation, afraid of death, any of those things. The space of alien minds that you can build is just so much faster than what evolution will give you. And with that also comes great responsibility for us to make sure that the kind of minds we create are the kind of minds that is good to create, minds that will share our values and be good for humanity and life, and also create minds that don't suffer.
Do you try to visualize the full space of alien minds that AI could be? Do you try to consider all the different kinds of intelligences, sort of generalizing what humans are able to do to the full spectrum of what intelligent creatures entities could do? I try, but I would say I fail.
I mean, it's very difficult for a human mind to really grapple with something so completely alien, maybe even for us. If we just try to imagine, how would it feel if we were completely indifferent towards death or individuality? Even if you just imagine that, for example,
you could just copy my knowledge of how to speak Swedish. Boom, now you can speak Swedish. And you could copy any of my cool experiences and then you could delete the ones you didn't like in your own life, just like that. It would already change quite a lot about how you feel as a human being, right? You probably spend less effort studying things if you just copy them and you might be less afraid of death because if the plane you're on starts to crash, you'd just be like, oh shucks, I haven't backed my brain up for four hours. So I'm going to lose all these wonderful experiences of this flight. We might also start feeling more compassionate maybe with other people if we can so readily share each other's experiences and our knowledge and feel more like a hive mind.
It's very hard though. I really feel very humble about this to grapple with it, how it might actually feel. The one thing which is so obvious though, which I think is just really worth reflecting on, is because the mind space of possible intelligence is so different from ours,

it's very dangerous if we assume they're going to be like us or anything like us.
Well, the entirety of human written history has been through poetry, through novels, been trying to describe through philosophy, trying to describe the human condition and what's entailed in it. Like Jessica said, fear of death and all those kinds of things, what is love, and all of that changes if you have a different kind of intelligence, like all of it. The entirety, all those poems, they're trying to sneak up to what the hell it means to be human, all of that changes. How AI concerns and existential crises that AI experiences, how that clashes with the human existential crisis, the human condition. It's hard to fathom, hard to predict.
It's hard, but it's fascinating to think about also. Even in the best case scenario where we don't lose control over the ever more powerful AI that we're building to other humans whose goals we think are horrible and where we don't lose control to the machines and AI provides the things that we want, even then you get into the questions you touched here. Maybe the struggle that it's actually hard to do things is part of the things that give this meaning as well.
For example, I found it so shocking that this new Microsoft GPT-4 commercial that they put together has this woman talking about showing this demo of how she's going to give a graduation speech to her beloved daughter and she asks GPT-4 to write it. If it's frigging 200 words or so, if I realized that my parents couldn't be bothered struggling a little bit to write 200 words and outsource that to their computer, I would feel really offended actually.
I wonder if eliminating too much of this struggle from our existence, do you think that would also take away a little bit of what means to be human?
We can't even predict. I had somebody mentioned to me that they started using chat GPT with a 3.5 and not 4.0 to write what they really feel to a person and they have a temper issue and they're basically trying to get chat GPT to rewrite it in a nicer way, to get the point across, but rewrite it in a nicer way. We're even removing the inner asshole from our communication. There's some positive aspects of that, but mostly it's just the transformation of how humans communicate. It's scary because so much of our society is based on this glue of communication and we're now using AI as the medium of communication that does the language for us. So much of the emotion that's laden in human communication, so much of the intent that's going to be handled by outsourced AI. How does that change everything?
How does that change the internal state of how we feel about other human beings? What makes us lonely? What makes us excited? What makes us afraid? How we fall in love? All that kind of stuff. For me personally, I have to confess the challenge is one of the things that really makes my life feel meaningful. If I go hike a mountain with my wife, I don't want to just press a button and be at the top. I want to struggle and come up there sweaty and feel, wow, we did this in the same way. I want to constantly work on myself to become a better person. If I say something in anger that I regret, I want to go back and really work on myself rather than just tell an AI from now on, always filter what I write, so I don't have to work on myself because then I'm not growing.
Yeah, but then again, it could be like with chess. An AI wants to significantly, obviously, supersedes the performance of humans. It will live in its own world and provide maybe a flourishing civilizations for humans, but we humans will continue hiking mountains and playing our games, even though AI is so much smarter, so much stronger, so much superior in every single way,

just like with chess. That's one possible hopeful trajectory here is that humans will continue to human, and AI will just be a medium that enables the human experience to flourish.
Yeah, I would phrase that as rebranding ourselves from homo sapiens to homo sentiens. Right now, sapiens, the ability to be intelligent, we've even put it in our species name. We're branding ourselves as the smartest information processing entity on the planet. That's clearly going to change if AI continues ahead, so maybe we should focus on the experience, instead the subjective experience that we have with homo sentiens, and say that's what's really valuable, the love, the connection, the other things, and get off our high horses and get rid of this hubris that only we can do integrals. So consciousness, the subjective experience is a fundamental value to what it means to be human. Make that the priority.
That feels like a hopeful direction to me, but that also requires more compassion, not just towards other humans, because they happen to be the smartest on the planet, but also towards all our other fellow creatures on this planet. And I personally feel right now, we're treating a lot of farm animals horribly, for example, and the excuse we're using is, oh, they're not as smart as us. But if we admit that we're not that smart in the grand scheme of things either in the post AI epoch, then surely we should value you the subjective experience of a cow also.
Well, allow me to briefly look at the book, which at this point is becoming more and more visionary than you've written, I guess, over five years ago, Life 3.0. So first of all, 3.0. What's 1.0? What's 2.0? What's 3.0? And how does that vision evolve? The vision in the book evolved to today. Life 1.0 is really dumb, like bacteria, and that it can't actually learn anything at all during the lifetime. The learning just comes from this genetic process from one generation to the next. Life 2.0 is us and other animals which have brains, which can learn during their lifetime a great deal. And you were born without being able to speak English. And at some point, you decided, hey, I want to upgrade my software. Let's install an English speaking module.
So you did. And Life 3.0 does not exist yet, can replace not only its software the way we can, but also its hardware. And that's where we're heading towards at high speed. We're already maybe 2.1 because we can put in an artificial knee, a pacemaker, et cetera, et cetera. And if Neuralink and other companies succeed, we'll be Life 2.2, et cetera. But the companies trying to build AGI are trying to make this, of course, full 3.0. And you can put that intelligence in something that also has no biological basis whatsoever.
So less constraints and more capabilities, just like the leap from 1.0 to 2.0. There is, nevertheless, you speaking so harshly about bacteria, so disrespectfully about bacteria. There is still the same kind of magic there that permeates Life 2.0 and 3.0. It seems like maybe the thing that's truly powerful about life, intelligence, and consciousness was already there in 1.0. Is it possible? I think we should be humble and not be so quick to make everything binary and say either it's there or it's not. Clearly, there's a great spectrum. And there is even a controversy about whether some unicellular organisms like amoebas can maybe learn a little bit after all. So apologies if I offended any bacteria here.
It wasn't my intent. It was more that I wanted to talk up how cool it is to actually have a brain where you can learn dramatically within your lifetime.
Typical human.
And the higher up you get from 1.0 to 2.0 to 3.0, the more you become the captain of your own

ship, the master of your own destiny, and the less you become a slave to whatever evolution gave you, right? By upgrading your software, we can be so different from previous generations and even from our parents, much more so than even a bacterium, no offense to them.
And if you can also swap out your hardware and take any physical form you want, of course, it's really disguised the limit.
Yeah, so it accelerates the rate at which you can perform the computation that determines your destiny. Yeah, and I think it's worth commenting a bit on what you mean in this context also if you swap things out a lot, right? This is controversial, but my current understanding is that life is best thought of not as a bag of meat or even a bag of elementary particles, but rather as a system which can process information and retain its own complexity, even though nature is always trying to mess it up. So it's all about information processing. And that makes it a lot like something like a wave in the ocean, which is not its water molecules, right? The water molecules bob up and down, but the wave moves forward, it's an information pattern. In the same way, you Lex, you're not the same Adams as during the first time you did with me, you've swapped out most of them, but it's still you. And the information pattern is still there. And if you could swap out your arms and whatever, you can still have this kind of continuity, it becomes much more sophisticated sort of wave forward in time where the information lives on. I lost both of my parents since our last podcast and it actually gives me a lot of solace that this way of thinking about them, they haven't entirely died because a lot of mommy and daddy's, sorry, I'm getting a little emotional here, but a lot of their values and ideas and even jokes and so on, they haven't gone away, right? Some of them live on, I can carry on some of them. And we also live on a lot of other and a lot of other people. So in this sense, even with life 2.0, we can to some extent, already transcend our physical bodies and our death. And particularly if you can share your own information, your own ideas with many others like you do in your podcast, then that's the closest immortality we can get with our bio bodies. You carry a little bit of them in you in some sense. Yeah. Do you miss them? You miss your mom and dad? Of course. Of course. What did
you learn about life from them if it can take a bit of a tangent? I know so many things. For starters, my fascination for math and the physical mysteries of our universe, I think I got a lot of that from my dad. But I think my obsession for really big questions and consciousness and so on that actually came mostly from my mom. And what I got from both of them, which is a very core part of really who I am, I think, is this feeling comfortable with not buying into what everybody else is saying, doing what I think is right. They both very much just did their own thing and sometimes they got flack for it and they did it anyway. That's why you've always been an inspiration to me, that you're at the top of your field and you're still willing to tackle the big questions in your own way. You're one of the people that represents MIT best to me. You've always been an inspiration to that. So it's good to hear that you got that from your mom and dad. Yeah, you're too kind. But yeah, I mean, the good reason to do science is because you're really curious, you want to figure out the truth. If you think this is how it is and everyone else says no, no, that's bullshit. And it's that way, you know, you stick with what you think is true. And even if everybody else keeps thinking it's bullshit, there's a certain, I always root for the underdog when I watch movies. And my dad once, one time, for

example, when I wrote one of my craziest papers ever, talking about our universe, ultimately being mathematical, which we're not going to get into today. I got this email from a quite famous professor saying this is not only bullshit, but it's going to ruin your career. You should stop doing this kind of stuff. I sent it to my dad. Do you know what he said? He replied with a quote from Dante. Segui il tuo corso e la sedire la gente. Follow your own path and let the people talk. Go dad. This is the kind of thing. He's dead, but that attitude is not.
How did losing them as a man, as a human being change you? How did it expand your thinking about the world? How did it expand your thinking about this thing we're talking about, which is humans creating another living sentient perhaps being? I think it mainly did two things. One of them just going through all their stuff after they had passed away and so on, just drove home to be how important it is to ask ourselves, why are we doing this things we do? Because it's inevitable that you look at some things they spent an enormous time on and you ask the, at hindsight, would they really have spent so much time on this or would they have done something that was more meaningful? So I've been looking more in my life now and asking, you know, why am I doing what I'm
doing? And I feel it should either be something I really enjoy doing or it should be something that I find really, really meaningful because it helps humanity. If it's none of those two categories, maybe I should spend less time on it. You know, the other thing is dealing with death up in person like this. It's actually made me less afraid of even less afraid of other people telling me that I'm an idiot, you know, which happens regularly and just let my life do my thing. And it made it a little bit easier for me to focus on what I feel is really important.
What about fear of your own death? Has it made it more real that this is something that happens? Yeah, it's made extremely real and I'm next in line in our family now, right? Me and my younger brother. But they both handled it with such dignity. That was a true inspiration also. They never complained about things and, you know, when you're old and your body starts falling apart, it's more and more to complain about. They looked at what could they still do that was meaningful. And they focused on that rather than wasting time talking about or even thinking much about things they were disappointed in. I think anyone can make themselves depressed
if they start their morning by making a list of grievances. Whereas if you start your day with a little meditation and just things you're grateful for, you basically choose to be a happy person. Because you only have a finite number of days. You should spend them. Make account. Being grateful. Yeah.
Well, you do happen to be working on a thing which seems to have potentially some of the greatest impact on human civilization of anything humans have ever created, which is artificial intelligence. This is on the both detailed technical level and in a high philosophical level you work on. So you've mentioned to me that there's an open letter that you're working on. It's actually going live in a few hours. So I've been having late nights and early mornings. It's been very exciting actually. In short, have you seen Don't Look Up? The film? Yes. Yes. I don't want to be the movie spoiler for anyone watching this who hasn't seen it. But if you're watching this, you haven't seen it, watch it. Because we are actually acting out. It's life imitating art. Humanity is doing exactly that right now, except it's an asteroid that we are building ourselves. Almost nobody is talking about it. People are squabbling

across the planet about all sorts of things which seem very minor compared to the asteroid that's about to hit us. Most politicians don't even have this on the radar. They think maybe in 100 years or whatever. Right now, we're at a fork on the road. This is the most important fork humanity has reached in its over 100,000 years on this planet. We're building effectively a new species that's smarter than us. It doesn't look so much like a species yet because it's mostly not embodied in robots, but that's the technicality which will soon be changed. This arrival of artificial general intelligence that can do all our jobs as well as us and probably shortly thereafter, superintelligence which greatly exceeds our cognitive abilities, it's going to either be the best thing ever to happen to humanity or the worst. I'm really quite confident that there is not that much middle ground there. But it would be fundamentally transformative to human civilization? Of course. Utterly and totally. Again, we branded ourselves as homo sapiens because it seemed like the basic thing. We're the king of the castle on this planet. We're the smart ones. If we can control everything else, this could very easily change. We're certainly not going to be the smartest on the planet very long if AI, unless AI progress just halts. We can talk more about why I think that's true because it's controversial. Then we can also talk about reasons you might think it's going to be the best thing ever and the reason you think it's going to be the end of humanity, which is of course super controversial. But what I think we can, anyone who's working on advanced AI can agree on is it's much like the film. Don't look up and that it's just really comical how little serious public debate there is about it, given how huge it is.
So what we're talking about is a development of currently things like GPT-4 and the signs it's showing of rapid improvement that may in the near term lead to development of super intelligent AGI, general AI systems and what kind of impact that has on society. When that thing achieves general
human level intelligence and then beyond that general super human level intelligence.
There's a lot of questions to explore here. One, you mentioned halt. Is that the content of the letter is to suggest that maybe we should pause the development of these systems?
Exactly. So this is very controversial. When we talked the first time, we talked about how I was involved in starting the Future Life Institute and we worked very hard on 2014-2015 was the mainstream AI safety. The idea that there even could be risks and that you could do things about them. Before then, a lot of people thought it was just really kooky to even talk about it and a lot of AI researchers felt worried that this was too flaky and could be bad for funding and that the people who talked about it just didn't understand AI. I'm very, very happy with how that's gone and that now it's completely mainstream. You're going to any AI conference and people talk
about AI safety and it's a nerdy technical field full of equations and blah blah.
As it should be. But there's this other thing which has been quite taboo up until now calling for slowdown. So what we've constantly been saying, including myself, I've been biting my tongue a lot, is that we don't need to slow down AI development. We just need to win this race, the wisdom race between the growing power of the AI and the growing wisdom with which we manage it. Rather than trying to slow down AI, let's just try to accelerate the wisdom. Do all this technical work to figure out how you can actually ensure that your powerful AI is going to do what you wanted to do and have society adapt also with incentives and regulations so that these things get put to good use. Sadly, that didn't pan out. The progress on technical AI

on capabilities has gone a lot faster than than many people thought back when we started this in 2014 turned out to be easier to build really advanced AI than we thought. On the other side, it's gone much slower than we hoped with getting policy makers and others to actually put incentives in place to steer this in the good direction. Maybe we should unpack it and talk a little bit about each. Why did it go faster than a lot of people thought? In hindsight, it's exactly like building flying machines. People spent a lot of time wondering about how do birds fly? That turned out to be really hard. Have you seen the TED Talk with a flying bird? Like a flying robotic bird? Yeah, flies around the audience, but it took 100 years longer to figure out how to do that than for the Wright brothers to build the first airplane because it turned out there was a much easier way to fly. Evolution picked a more complicated one because it had its hands tied. It could only build a machine that could assemble itself, which the Wright brothers didn't care about. They can only build a machine that used only the most common atoms in the periodic table. Wright brothers didn't care about that. They could use steel, iron, atoms. It had to be built to repair itself and it also had to be incredibly fuel efficient. A lot of birds use less than half the fuel of a remote control plane flying the same distance. For humans, just throw a little more, put a little more fuel in a roof. There you go, 100 years earlier. That's exactly what's happening now with these large language models. The brain is incredibly complicated. Many people made the mistake. You're thinking we have
to figure out how the brain does human level AI first before we could build in a machine. That was completely wrong. You can take an incredibly simple computational system called a transformer network and just train it to do something incredibly dumb. Just read a gigantic amount of text and try to predict the next word. It turns out if you just throw a ton of compute at that and a ton of data, it gets to be frighteningly good, like GPT-4, which I've been playing with so much since it came out. There's still some debate about whether that can get you all the way to full human level or not. We can come back to the details of that and how you might get the human level AI even if large language models don't. Can you briefly, if it's just a small tangent comment on your feelings about GPT-4, suggest that you're impressed by this rate of progress, but where is it? Can GPT-4 reason? What are the intuitions? What are human interpretable words you can assign to the capabilities of GPT-4 that makes you so damn impressed with it? I'm both very excited about it and terrified. It's an interesting mixture of emotions.
All the best things in life include those two somehow.
Yeah, I can absolutely reason. Anyone who hasn't played with it, I highly recommend doing that before dissing it. It can do quite remarkable reasoning. I've had to do a lot of things, which I realized I couldn't do that myself that well even. It obviously does it dramatically faster than we do too when you watch a type. It's doing that while servicing a massive number of other humans at the same time. At the same time, it cannot reason as well as a human can
on some tasks. It's obviously a limitation from its architecture. We have in our heads what in GeekSpeak is called a recurrent neural network. There are loops. Information can go from this neuron to this neuron to this neuron and then back to this one. You can ruminate on something for a while. You can self-reflect a lot. These large language models, they cannot. It's a so-called transformer where it's just like a one-way street of information basically. In GeekSpeak,

it's called a feed-forward neural network. It's only so deep. It can only do logic that's that many steps and that deep. You can create problems which will fail to solve for that reason. The fact that it can do so amazing things with this incredibly simple architecture already is quite stunning. What we see in my lab at MIT when we look inside large language models to try to figure out how they're doing it, that's the key core focus of our research. It's called mechanistic interpretability in GeekSpeak. You have this machine that does something smart. You try to reverse engineer. See how does it do it? I think of it also as artificial neuroscience. That's exactly what neuroscientists do with actual brains. But here you have the advantage that you don't have to worry about measurement errors. You can see what every neuron is doing all the time. A recurrent thing we see again and again, there's been a number of beautiful papers quite recently by a lot of researchers. Some of them here, even in this area, is where when they figure out how something is done, you can say, oh, man, that's such a dumb way of doing it. You immediately see how it can be improved. For example, there was a beautiful paper recently where they figured out how a large language model stores certain facts like Eiffel Tower is in Paris. They figured out exactly how it's stored and the proof that they understood it was they could edit it. They changed some of the synapses in it and then they asked it, where's the Eiffel Tower? And they said, it's in Rome. And then they asked you, how do you get there? Oh, how do you get there from Germany? Oh, you take this train and the Roma Termini train station and this and that.

And what might you see if you're in front of it? Oh, you might see the Colosseum.

So they had edited it. So they literally moved it to Rome. But the way that it's storing this information, it's incredibly dumb for any fellow nerds listening to this.

There was a big matrix and roughly speaking, there are certain row and column vectors which encode these things and they correspond very hand-wavely to principal components. And it would be much more efficient for a sparse matrix to store in the database. And everything, so far we've figured out how these things do are ways where you can see they can easily be improved. And the fact that this particular architecture has some roadblocks built into it is in no way going to prevent crafty researchers from quickly finding workarounds and making other kinds of architectures go all the way. So in short, it's turned out to be a lot easier to build close to human intelligence than we thought. And that means our runway as a species to get our shit together has shortened. And it seems like the scary thing about the effectiveness of large language models. So Sam Altman, I recently had a conversation with, and he really showed that

the leap from GPT-3 to GPT-4 has to do with just a bunch of hacks, a bunch of little explorations with smart researchers doing a few little fixes here and there. It's not some fundamental leap and transformation in the architecture. And more data and more compute.

And more data and compute, but he said the big leaps has to do with not the data and the compute, but just learning this new discipline, just like you said. So researchers are going to look at these architectures and there might be big leaps where you realize, wait, why are we doing this in this dumb way? And all of a sudden this model is 10x smarter. And that can happen on any one day, on any one Tuesday or Wednesday afternoon. And then all of a sudden you have a system that's

10x smarter. It seems like it's such a new discipline. It's such a new, like we understand

so little about why this thing works so damn well, that the linear improvement of compute, or exponential, but the steady improvement of compute, steady improvement of the data may not be the thing that even leads to the next leap. It could be a surprise little hack that improves everything. Or a lot of little leaps here and there because, because so much of this is out on the open also. So many smart people are looking at this and trying to figure out little leaps here and there. And it becomes this sort of collective race where if people, a lot of people feel if I don't take the leap, someone else will. And it's actually very crucial for the other part of it. Why do we want to slow this down? So again, what this open letter is calling for is just pausing all training of systems that are more powerful than GPT for for six months. Give a chance for the labs to coordinate a bit on safety and for society to adapt, give the right incentives to the labs. Because I, you know, you've interviewed a lot of these people who lead these labs. And you know, just as well as I do, they're good people. They're idealistic people. They're doing this first and foremost because they believe that AI has a huge potential to help humanity. And but at the same time, they are trapped in this horrible race to the bottom.
Have you read Meditations on Moloch by Scott Alexander? Yes. Yeah, it's a beautiful essay on this poem by Ginsburg, where he interprets it as being about this monster. It's this game theory monster that that pits people into against each other in this race, the bottom where everybody ultimately loses the edit. The evil thing about this monster is even though everybody sees it and understands, they still can't get out of the race, right? Most a good fraction of all the bad things that we humans do are caused by Moloch. And I like Scott Alexander's naming of the monster. So we can, we humans can think of it as an F a thing. If you look at why do we have overfishing? Why do we have more generally the tragedy of the commons? Why is it that
to live or a I don't know if you had her on your podcast. Yeah, she's become a friend. Yeah. Great. She made this awesome point recently that beauty filters that a lot of female influencers feel pressure to use are exactly Moloch in action again. First, nobody was using them. And people saw them just the way they were. And then some of them started using it and becoming ever more plastic fantastic. And then the other ones that weren't using it started to realize that if they want to just keep their, their market share, they have to start using it too. And then you're in a situation where they're all using it. And none of them has any more market share or less than before. So nobody gained anything, everybody lost. And they have to keep becoming ever more plastic fantastic also. Right. And but nobody can go back to the old way because
it's just too costly, right? Moloch is everywhere. And Moloch is not a new arrival on the scene either. We humans have developed a lot of collaboration mechanisms to help us fight back against Moloch through various kinds of constructive collaboration. The Soviet Union and the United States did sign the number of arms control treaties against Moloch who is trying to stoke them into unnecessarily risky nuclear arms races, etc, etc. And this is exactly what's happening on the AI front.
This time, it's a little bit geopolitics, but it's mostly money, where there's just so much commercial pressure. You know, if you take any of these leaders of the top tech companies, and if they just say, you know, this is too risky, I want to pause for six months, they're going to get a lot of pressure from shareholders and others are like, well, you know, if you pause, but those guys don't pause, we're, we don't want to get our lunch eaten. Yeah. And shareholders even have the power to replace the executives in the worst case, right? So we did this open letter

because we want to help these idealistic tech executives to do what their heart tells them by providing enough public pressure on the whole sector to just pause so that they can all pause in a coordinated fashion. And I think without the public pressure, none of them can do it alone, push back against their shareholders, no matter how good-hearted they are. Moloch is a really powerful foe. So the idea is to, for the major developers of AI systems like this, so we're talking about Microsoft, Google, Meta, and anyone else? OpenAI is very close with Microsoft now, of course, and there are plenty of smaller players. For example, Anthrope is very impressive, there's Conjecture, there's many, many, many players. I don't want to make a long list to leave anyone out. And for that reason, it's so important that some coordination happens, that there's external pressure on all of them, saying you all need to pause. Because then the people, the researchers in these organizations, the leaders who want to slow down a little bit, they can say their shareholders, you know, everybody's slowing down because of this pressure, and it's the right thing to do. Have you seen in history their examples where it's possible to pause the Moloch? Absolutely. And even like human cloning, for example, you could make so much money
on human cloning. Why aren't we doing it? Because biologists thought hard about this and felt like this is way too risky. They got together in the 70s in Asilomar and decided even
to stop a lot more stuff, also just editing the human germline, gene editing that goes into our offspring and decided let's not do this because it's too unpredictable what it's going to lead to. We could lose control over what happens to our species. So they paused.
There was a ton of money to be made there. So it's very doable, but you just need a public awareness of what the risks are and the broader community coming in and saying, hey, let's slow down. And another common pushback I get today is we can't stop in the West because China, and in China, undoubtedly, they also get told we can't slow down because the West, because both sides think they're the good guy. But look at human cloning.
Did China forge ahead with human cloning? There's been exactly one human cloning that's actually been done that I know of. It was done by a Chinese guy. Do you know where he is now? In jail. And who put him there? Who? Chinese government. Not because Westerners said China allowed this. No, the Chinese government put him there because they also felt they liked control, the Chinese government. If anything, maybe they are even more concerned about having control than Western governments have no incentive of just losing control over where everything is going. And you can also see the Ernie bot that was released by I believe Baidu recently. They got a lot of pushback from the government and had to reign it in in a big way.
I think once this basic message comes out that this isn't an arms race, it's a suicide race, where everybody loses if anybody's AI goes out of control. It really changes the whole dynamic.
I'll say this again, because this is a very basic point I think a lot of people get wrong.
Because a lot of people dismiss the whole idea that AI can really get very superhuman, because they think there's something really magical about intelligence such that it can only exist in human minds. Because they believe that, they think it's kind of get to just more or less GPT-4++ and then that's it. They don't see it as a suicide race. They think whoever gets that first, they're going to control the world, they're going to win. That's not how it's going to be. And we can talk again about the scientific arguments from why it's not going to stop there. But the way it's going to be is if anybody completely loses control and you don't care,

if someone manages to take over the world who really doesn't share your goals, you probably don't really even care very much about what nationality they have. You're not going to like it, much worse than today. If you live in Orwellia in dystopia, what do you care who created it, right? And if it goes farther and we just lose control even to the machines so that it's not us versus them, it's us versus it, what do you care who created this unaligned entity which has goals different from humans ultimately and we get marginalized, we get made obsolete, we get replaced. That's why what I mean when I say it's a suicide race. It's kind of like we're rushing towards this cliff. But the closer the cliff we get, the more scenic the views are and the more money there is there and so we keep going. But we have to also stop at some point, right? Quit while we're ahead. And it's a suicide race which cannot be won. But the way to really benefit from it is to continue developing awesome AI a little bit slower so we make it safe, make sure it does the things that humans want and create a condition where everybody wins. And that technology has shown us that geopolitics and politics in general is not a zero sum game at all. So there is some rate of development that will lead us as a human species to lose control of this thing. And the hope you have is that there's some lower level of development which will not allow us to lose control. This is an interesting thought you have about losing control. So if you are somebody like Sander Prachai or Sam Altman at the head of a company like this, you're saying if they develop an AGI, they too will lose control of it. So no one person can maintain control. No group of individuals can maintain control. If it's created very, very soon and is a big black box that we don't understand like the large language models, yeah, then I'm very confident they're going to lose control. But this isn't just me saying it. You know, Sam Altman and them as the Sabbaths have both said themselves, acknowledge that there's really great risks with this and they want to slow down once they feel it gets scary. But it's clear that they're stuck in this. Again, Malak is forcing them to go a little faster than they're comfortable with because of pressure from just commercial pressures, right? To get a bit optimistic here, of course this is a problem that can be ultimately solved. To win this wisdom race, it's clear that what we hope that is going to happen hasn't happened. The capability progress has gone faster than a lot of people thought and the progress in the public sphere of policymaking and so on has gone slower than we thought. Even the technical AI safety has gone slower. A lot of the technical safety research was kind of banking on that large language models and other poorly understood systems couldn't get us all the way. But you had to build more of a kind of intelligence that you could understand. Maybe it could prove itself safe, you know, things like this. And I'm quite confident that this can be done so we can reap all the benefits. But we cannot do it as quickly as this out of control express train we are on now is going to get the AGI. That's why we need a little more time, I feel. Is there something to be said with like Sam Allman talked about which is while we're in the pre-AGI stage to release often and as transparently as possible to learn a lot. So as opposed to being extremely cautious, release a lot. Don't invest in a closed development where you focus on AI safety while it's somewhat dumb, quote unquote, release as often as possible. And as you start to see signs of human level intelligence or super human level intelligence, then you put a halt on it. Well, what a lot of safety researchers have been saying for many years is that the most dangerous things you can do with an AI is first of all, teach it to write code.

Yeah, because that's the first step towards recursive self-improvement which can take it from AGI to much higher levels. Okay, oops, we've done that. And another thing, high risk is connected to the internet. Let it go to websites, download stuff on its own and talk to people. Oops, we've done that already. You know, Elias Jukowski, you said you interviewed him recently, right? Yes, yes. So he had this tweet recently which gave me one of the best laughs in a while where he was like, hey, people used to make fun of me and say you're so stupid, Elias, because you're saying you're saying you have to worry. Obviously, developers, once they get to like really strong AI, first thing you're going to do is like never connect it to the internet, keep it in the box where, you know, you can really study it safe. So he had written it in the like in the meme form, so it's like then and then that and now. Let's LOL, let's make a chatbot.
And the third thing is Stuart Russell, you know, amazing AI researcher. He has argued for a while that we should never teach AI anything about humans. Above all, we should never let it learn about human psychology and how you manipulate humans. That's the most dangerous kind of knowledge
you can give it. Yeah, you can teach it all it needs to know about how to cure cancer and stuff like that, but don't let it read Daniel Kahneman's book about cognitive biases and all that. And then oops, LOL, you know, let's invent social media. I'll recommend our algorithms which do exactly that. They get so good at knowing us and pressing our buttons that we're starting to create a world now where we're just having ever more hatred because they figured out that these algorithms, not for out of evil, but just to make money on advertising that the best way to get more engagement to euphemism, get people glued to their little rectangles, right, is just to make them pissed off. That's really interesting that a large AI system that's doing the recommender system kind of task on social media is basically just studying human beings because it's a bunch of us rats giving it signal, nonstop signal. It'll show a thing and then we give signal and whether we spread that thing, we like that thing, that thing increases our engagement, gets us to return to the platform. It has that on the scale of hundreds of millions of people constantly. So it's just learning and learning and learning and presumably if the parameter, the number of parameters in neural network that's doing the learning and more and to end the learning is the more it's able to just basically encode how to manipulate human behavior, how to control humans at scale. Exactly and that is not something I think is in humanity's interest. Yes. Right now it's mainly letting some humans manipulate other humans for profit and power which already
caused a lot of damage and eventually that's a sort of skill that can make AIs persuade humans to let them escape whatever safety precautions we put. There was a really nice article in the New York Times recently by Yuval Noah Harari and two co-authors including Tristan Harris from the Social Dilemma and they have this phrase in there I love. Humanity's first contact with advanced AI was social media and we lost that one. We now live in a country where there's much more hate in the world where there's much more hate in fact and in our democracy that
we're having this conversation and people can't even agree on who won the last election you know and we humans often point fingers at other humans and say it's their fault but it's really malloc and these AI algorithms. We got the algorithms and then malloc pitted the social media companies against each other so nobody could have a less creepy algorithm because then they would lose out on revenue to the other company. Is there any way to win that battle back

just if we just linger on this one battle that we've lost in terms of social media? Is it possible to redesign social media this very medium in which we use as a civilization to communicate with each other to have these kinds of conversations to have discourse to try to figure out how to solve the biggest problems in the world whether that's nuclear war or the development of AGI. Is it possible to do social media correctly? I think it's not only possible but it's necessary. Who are we kidding that we're going to be able to solve all these other challenges if we can't even have a conversation with each other that's constructive. The whole idea the key idea of democracy is that you get a bunch of people together and they have a real conversation the ones you try to foster on this podcast or you respectfully listen to people you disagree with and you realize actually you know there are some things actually we some common ground we have and
we both agree let's not have nuclear wars let's not do that etc etc. We're kidding ourselves thinking we can face off the second contact with with ever more powerful AI that's happening now with these large language models if we can't even have a functional conversation in the public space that's why I started the improve the news project improve the news.org but I'm an optimist fundamentally in that there is a lot of intrinsic goodness in people and that what makes the difference between someone doing good things for humanity and bad things is not some sort of fairy tale thing that this person was born with an evil gene and this one was not born with a good gene. No I think it's whether we put whether people find themselves in situations that bring out the best in them or the bring out the worst in them and I feel we're building an internet and a society that brings out the worst. But it doesn't have to be that way. No it does not. It's possible to create incentives and also create incentives that make money that both make money and bring out the best in people. I mean in the long term it's not a good investment for anyone to have a nuclear war for example and is it a good investment for humanity if we just ultimately replace all humans by machines and then are so obsolete that eventually there's no humans
left. Well it depends again somehow you do the math but I would say by any reasonable economic standard if you look at the future income of humans and there aren't any that's not a good investment
moreover like why can't we have a little bit of pride in our species dammit you know why should we just build another species that gets rid of us. If we were Neanderthals would we really consider it a smart move if we had really advanced biotech to build homo sapiens? You know you might say hey
Max you know yeah let's build these homo sapiens they're gonna be smarter than us maybe they can help us defend this better against predators and help fix up our caves make them nicer and we'll control them undoubtedly you know so then they build a couple a little baby girl a little baby boy you know and and then you have some some wise old Neanderthal elders like hmm I'm scared that we're opening a Pandora's box here and that we're gonna get outsmarted by these super Neanderthal intelligences and there won't be any Neanderthals left and then but then you have a bunch of others in the cave right you are you such a luddite scaremonger of course they're gonna want to keep us around because we are their creators and and why you know the smarter I think
the smarter they get the nicer they're gonna get they're gonna leave us they're gonna they're gonna

want us around and it's gonna be fine and and besides look at these babies they're so cute it's clearly they're totally harmless that's exact those babies are exactly GPT-4 yeah it's not I want to be clear it's not GPT-4 that's terrifying it's the GPT-4 is a baby technology you know and Microsoft even had a paper recently out with a title something like sparkles of AGI whatever basically saying this is baby AI like these little Neanderthal babies and it's gonna grow up there's gonna be other systems from from the same company from other companies will be way more powerful and but they're gonna take all the things ideas from these babies
and before we know it we're gonna be like those last Neanderthals who are pretty disappointed and when they realized that they were getting replaced well this interesting point you make which is the programming it's it's not really possible that GPT-4 is already the kind of system that can change everything by writing programs so it's yeah it's because it's life 2.0 the systems I'm afraid of are gonna look nothing like a large language model and they're not gonna but once it gets once it or other people figure out a way of using this tech to make much better tech right it's just constantly replacing its software and from everything we've seen about how how these work under the hood they're like the minimum viable intelligence they do everything in a dumbest way that still works sort of yeah and so they are life 3.0 except when they replace their software it's a lot faster than when you when you decide to learn Swedish and moreover they think a lot faster than us too so when you know we don't think on how one logical step every nanosecond or few or so the way they do and we can't also just suddenly scale up our hardware massively in the cloud so limited right so they are they are also life consume become a little bit more like life 3.0 in that if they need more hardware hey just rent it in the cloud you know how do you pay for it well with all the services you provide
and what we haven't seen yet which could change a lot is a entire software system so right now programming is done sort of in bits and pieces as as an assistant tool to humans but I do a lot of programming and with the kind of stuff that GPT-4 is able to do I mean is replacing a lot what I'm able to do but I you still need a human in the loop to kind of manage the design of things manage like what are the prompts that generate the kind of stuff to do some basic adjustment of the code to do some debugging but if it's possible to add on top of GPT-4 kind of feedback loop of of of self-debugging improving the code and then you launch that system onto the wild on the internet because everything is connected and have it do things have it interact with humans and then get that feedback now you have this giant ecosystem of humans this is one of the things that Elon Musk recently sort of tweeted as a case why everyone needs to pay seven dollars or whatever for twitter to make sure they're real they make sure they're real we're now going to be living in a world where the the bots are getting smarter and smarter and smarter to a degree where where you can't you can't tell the difference between a human and a bot that's right and now you can have bots outnumber humans by one million to one which is why he's making a case why you have to pay to prove you're human which is one of the only mechanisms to prove which is depressing and I yeah I feel we have to remember as individuals we should from time to time ask ourselves why are we doing what we're doing all right and as a species we need to do that too so if we're building as you say machines that are outnumbering us and more and more outsmarting us and replacing us on the job market not just for the dangerous and and boring tasks

but also for writing poems and doing art and things that a lot of people find really meaningful
gotta ask ourselves why why are we doing this we are the answer is malloc is tricking us into
doing it and it's such a clever trick that even though we see the trick we still have no choice
but to fall for it right come also the thing you said about you using uh co-pilot AI tools to
program faster how many times what factor faster would you say you code now does it go twice as
fast or I don't really uh because it's a new tool yeah it's I don't know if speed is significantly
improved but it feels like I'm a year away from being uh five to ten times faster so if that's
typical for programmers then uh you're already seeing another kind of self recursive self-
improvement
right because previously one like a major generation of improvement of the codes would
happen on the human r&d timescale and now if that's five times shorter then it's going to take five
times less time than otherwise would to develop the next level of these tools and so on so this
these these are the this is exactly the sort of beginning of an of an intelligence explosion
there can be humans in the loop a lot in the early stages and then eventually
humans are needed less and less and the machines can more kind of go alone but you
what you weren't you said there is just the exact example of these sort of things another thing which
which um I was kind of lying on my psychiatrist imagining I'm on a psychiatrist's couch here
saying what are my fears that people would do with AI systems another so I mentioned three
that I had fears about many years ago that they would do namely uh teach you the code
connected to the internet and teach it to manipulate humans a fourth one is building an API
where code can control the super powerful thing right that's very unfortunate because
one thing that systems like GPT-4 have going for them is that they are an oracle in the sense that
they just answer questions there is no robot connected to GPT-4 GPT-4 can't go and do stock
trading based on its thinking yeah it's not an agent an intelligent agent is something that
takes in information from the world processes it to figure out what action to take based on its
goals that it has and then does something back on the world but once you have an API for example
GPT-4 nothing stops Joe Schmo and a lot of other people from building real agents which just keep
making calls somewhere in some inner loop somewhere to these powerful oracle systems
which makes them themselves much more powerful that's another kind of unfortunate development
which I think we would have been better off delaying I don't want to pick on any particular
companies I think they're all under a lot of pressure to make money yeah and again we the
reason we're calling for this pause is to give them all cover to do what they know is the right
thing slow down a little bit at this point but everything we've talked about I hope we'll can
we'll make it clear to people watching this you know why these sort of human level tools can cause
a gradual acceleration you keep using yesterday's technology to build tomorrow's technology yeah
and
when you do that over and over again you naturally get an explosion you know that's the definition
of an explosion in science right like if you have two people they fall in love now you have four
people and then they can make more babies and now you have eight people and then then you have
1632
64 etc that's we call that a population explosion where it's just that each if it's instead free
neutrons in a nuclear reaction that if each one can make more than one then you get an exponential
growth in that we call it a nuclear explosion all explosions are like that in an intelligence

explosion it's just exactly the same principle that some quantities some amount of intelligence can make more intelligence than that and then repeat you always get the exponentials what's your intuition why does you mention there's some technical reasons why it doesn't stop at a certain point what's your intuition and do you have any intuition why it might stop
it's obviously going to stop when it bumps up against the laws of physics
there are some things you just can't do no matter how smart you are allegedly
and because we don't know the full laws of physics yeah right Seth Lloyd wrote a really cool paper on the physical limits on computation for example if you make it put too much energy into it and the finite space it'll turn into a black hole you can't move information around fast and the speed of light stuff like that but it's hard to store way more than than a modest number of bits per atom etc but you know those limits are just astronomically above like 30 orders of magnitude
about where we are now so bigger different bigger jump in intelligence than if you go from a from an ant to a human I think of course what we want to do is have have a controlled thing a nuclear reactor you put moderators in to make sure exactly it doesn't blow up out of control right when we do experiments with biology and cells and so on you know we also try to make sure it doesn't get out of control and we can do this with AI too the thing is we haven't succeeded yet and malloc is exactly doing the opposite just fueling just egging everybody on faster faster faster or the other company is going to catch up with you or the other country is going to catch up with you we do this we have to want this stuff we have and I don't believe in this just asking people to look into their hearts and do the right thing it's easier for others to say that but like if if you're in the situation where your company is going to get screwed if you by other companies they're not stopping you know you're putting people in a very hard situation the right thing to do is change the whole incentive structure instead and this is not an old maybe I should say one more thing about this because malloc has been around as humanity's number one or number two enemy since the beginning of civilization and we came up with some really cool counter measures like first of all already over a hundred thousand years ago evolution realized that it was a very unhelpful that people kept killing each other all the time so it genetically gave us compassion and made it so that like if you get two drunk dudes getting into a pointless bar fight they might give each other black eyes but they have a lot of inhibition towards just killing each other that's a and similarly if you find a baby lying on the street when you go out for your morning jog tomorrow you're gonna stop and pick it up right even though it may be a make you late for your next podcast so evolution gave us these genes that make our own egoistic incentives
more aligned with what's good for the greater group or part of right and then as we got a bit more sophisticated and developed language we invented gossip which is also a fantastic anti malloc right because now it it's really discourages liars moochers cheaters because it their own incentive now is not to do this because word quickly gets around and then suddenly people aren't going to invite them to their dinners anymore and or trust them and then when we got still more sophisticated and bigger societies you know invented the legal system where even strangers
who didn't couldn't rely on gossip and things like this would treat each other would have an incentive

now those guys in the bar fights even if they someone is so drunk that he actually wants to kill the other guy he also has a little thought in the back of his head that you know do i really want to spend the next 10 years eating like really crappy food in a small room i'm just gonna i'm just gonna chill out you know so and we we similarly have tried to give these incentives to our corporations by having having regulation and all sorts of oversight so that their incentives are aligned with the greater good we tried really hard and the big problem that we're failing now is not that we haven't tried before but it's just that the tech is growing much is developing much faster than the regulators been able to keep up right so regulators it's kind of comical that european union right now is doing this AI act right which and in the beginning they had a little opt-out exception that gpt4 would be completely excluded from regulation brilliant idea what's the logic behind that some lobbyists pushed successfully for this so we were actually quite involved with the future life institute mark brackell mr uke anthony agir and others you know we're quite involved with talking to very educating various people involved in this process about these general purpose AI models coming and pointing out that they would become the laughing stock if they didn't put it in so it the french started pushing for it got put in to the draft and it looked like all was good and then there was a huge counter push from lobbyists yeah there were more lobbyists in brussels from tech companies and from oil companies for example and it looked like it might is we're going to maybe get taken out again and now gpt4 happened and i think it's going to stay in but this just shows you know malloc can be defeated but the the challenge we're facing is that the tech
is generally much faster than what the policy makers are and a lot of the policy makers also don't have a tech background so it's you know we really need to work hard to educate them on on how on what's taking place here so so we're getting the situation where the first kind of non so you know i define artificial intelligence just as non biological intelligence all right and by that definition a company a corporation is also an artificial intelligence because the corporation isn't it's humans it's a system if its CEO decides the CEO of a tobacco company decides one morning the CEO he doesn't want to sell cigarettes anymore they'll just put another CEO in
there it's not enough to align the incentives of individual people or align individual computers incentives to their owners which is what technically iSafety research is about you also have to align the incentives of corporations with a greater good and some corporations have gotten so big and so powerful very quickly that in many cases their lobbyists instead align the regulators to what they want rather than the other way around it's a classic regulatory capture all right is is the thing that the slowdown hopes to achieve is give enough time to the regulators to catch up or enough
time to the companies themselves to breathe and understand how to do AI safety correctly i think both and but i think that the vision path to success i see is first you give a breather actually to to the people in these companies their leadership who wants to do the right thing and they all have safety teams and so on on their companies give them a chance to get together with the other companies and the outside pressure can also help catalyze that right and and work out
what is it that's what are the reasonable safety requirements one should put on future systems

before they get rolled out there are a lot of people also in academia and elsewhere outside of these companies who can be brought into this and have a lot of very good ideas and then i think it's very realistic that within six months you can get these people coming up so here's a white paper here's where we all think is reasonable um you know you didn't just because cars killed a lot of people you didn't ban cars but they got together a bunch of people and decided you know in order to be allowed to sell a car it has to have a seat belt in it there the analogous things that you can start requiring a future AI systems so that they are are safe and uh once this have this heavy heavy lifting this intellectual work has been done by experts in the field which can be done quickly i think it's going to be quite easy to get policymakers to to see yeah this is a good idea and it's it's you know for the fight for the companies to fight malloc they want and i believe sam altman has explicitly called for this they want the regulators to actually adopt it so that their competition is going to abide by it too right you don't want uh you don't want to be enacting all these principles then you abide by them and then there's this one little company that doesn't sign on to it and then now they can gradually overtake you then the companies will get be able to sleep secure knowing that everybody's playing by the same rules
so do you think it's possible to develop guardrails that keep the systems from from basically damaging irreparably humanity while still enabling sort of the capitalist fueled competition between companies as they develop how to best make money with this AI you think there's a balancing that's possible absolutely i mean we've seen that in many other sectors where you've had the free market produce quite good things without causing particular harm um when the guardrails are there and they work you know capitalism is a very good way of optimizing for just getting the same things on more efficiently it was but it was good you know and like in hindsight i've never met anyone even even on parties way over on the right in in any country who think it was a bad it thinks it was a terrible idea to ban child labor for example yeah but it seems like this particular technology has gotten so good so fast become powerful to a degree where you could see in the near term the ability to make a lot of money and to put guardrails develop guardrails quickly in that kind of context seems to be tricky it's not similar to cars or child labor it seems like the opportunity to make a lot of money here very quickly is right here yeah again it's there's this cliff yeah this gets quite seen in the closer the cliff there you go more the more there more money there is more gold in gets there on the ground you can pick up or whatever it's you want to drive there very fast but it's not in anyone's incentive that we go over the cliff and it's not like everybody's in their own car all the cars are connected together with a chain yeah so if anyone goes over they'll start dragging others down the others down too and so ultimately it's in the selfish interests also of the people in the companies to slow down when the when you start seeing the contours of the cliff there in front of you right and the problem is that even though the people who are building the technology and the CEOs they really get it the shareholders and these other market forces they are people who don't honestly understand that the cliff is there they usually don't you have to get quite into the weeds to really appreciate how powerful this is and how fast and a lot of people are even still stuck again in this idea that intelligence in this carbon chauvinism as I like to call it that you can only have our level of intelligence in humans that there's something magical about it whereas the people in the tech companies who build this stuff they all realize that intelligence is information processing

of a certain kind and it really doesn't matter at all whether the information is processed by carbon atoms in neurons and brains or by silicon atoms and some technology we build so you brought up capitalism earlier and there are a lot of people who love capitalism and a lot of people who really really don't and it struck me recently that what's happening with capitalism here is exactly analogous to the way in which super intelligence might wipe us out so you know I studied economics for my undergrad stock in school economics yay well no no I tell me so I was very interested in how how you could use market forces to just get stuff done more efficiently but give the right incentives to market so that it wouldn't do really bad things so Dylan had Phil Manel who's a professor and colleague of mine at MIT wrote this really interesting paper with some collaborators recently where they proved mathematically that if you just take one goal that you just optimize for on and on and on indefinitely that you think is gonna bring you in the right direction but basically always happens is in the beginning it will make things better for you but if you keep going at some point it's going to start making things worse for you again and then gradually it's going to make it really really terrible so just as a simple the way I think of the proof is like suppose you want to go from here back to Austin for example and you're like okay yeah let's just let's go south but you put in exactly the right sort of the right direction just optimize that south as possible you get closer and closer to Austin but uh you there's always some little error so you you're not going exactly towards Austin but you get pretty close but eventually you start going away again and eventually you're gonna be leaving the solar system yeah and they they proved it's beautiful mathematical proof this happens generally and this is very important for AI because for even though Stuart Russell has written a book and given a lot of talks on why it's a bad idea to have AI just blindly optimize something that's what pretty much all our systems do yeah we have something called the loss function that we're just minimizing or reward function we're just minimize maximizing and um capitalism is exactly like that too we want we wanted to get stuff done more efficiently that people wanted so introduce the free market things got done much more efficiently
than they did and and say communism right and it got better but then it just kept optimizing it and kept optimizing and you got ever bigger companies and ever more efficient information processing and now also very much powered by it and eventually a lot of people are beginning to feel wait we're kind of optimizing a bit too much like why did we just chop down half the rain for us you know and why why did suddenly these regulators get captured by lobbyists and so on it's just the same optimization that's been running for too long if you have an AI that actually has power over the world and you just give it one goal and just like keep optimizing that most likely everybody's gonna be like yay this is great in the beginning things are getting better but um it's almost impossible to give it exactly the right direction to optimize in and then eventually all hey breaks loose right nick boss drum and others are giving it examples that sound quite silly like what if you just want to like tell it to cure cancer or something and that's all you tell it maybe it's going to decide to take over entire confidence just so we can get more super computer facilities in there and figure out a cure cancer backwards and then you're like wait that's not what I wanted right and the the the the issue with capitalism and the issue with running away I have kind of merged now because the malloc I talked about is exactly the capitalist malloc that we have built an economy that has is optimized for only one

thing profit right and that worked great back when things were very inefficient and then now it's getting done better and it worked great as long as the companies were small enough that they couldn't
capture the regulators but that's not true anymore but they keep optimizing and now we
they realize that that they can these companies can make even more profit by building ever more powerful AI even if it's reckless but optimize more more more more more so this is malloc again showing up and I just want to anyone here who has any concerns about about late-stage capitalism having gone a little too far you should worry about superintelligence because it's the same
villain in both cases it's it's malloc and optimizing one objective function aggressively
blindly is going to take us there yeah we have this pause from time to time and look into our hearts and that's why are we doing this is this am I still going towards Austin or have I gone too far you know maybe we should change direction and that is the idea behind the halt for six months why six months that seems like a very short period just can we just linger and explore
different ideas here because this feels like a really important moment in human history
where pausing would actually have a significant positive effect we said six months because
we figured the number one pushback we were going to get in the west was like but China
and everybody knows there's no way that China is going to catch up with the west on this in six months so it's that argument goes off the table and you can forget about geopolitical competition and just focus on the real issue that's why we put this that's really interesting
but you've already made the case that even for China if you actually want to take on that argument China too would not be bothered by a longer halt because they don't want to lose control
even more than the west doesn't that's what I think that's a really interesting argument
like I have to actually really think about that which the the kind of thing people assume is if you develop an AGI that open AI if they're the ones that do it for example they're going to win
but you're saying no they're everybody loses yeah it's going to get better and better and better and then kaboom we all lose that's what's going to happen when losing win a defined on a metric of basically quality of life for human civilization and for sam altman to be people on my personal guess you know and people can quibble with this is that we're just gonna there won't be any humans that's it that's what I mean by lose you know if you we can see in history once you have some species or some group of people who aren't needed anymore doesn't usually work out so well for them
right yeah there were a lot of horses for the way used for traffic in boston and then the
car got invented and most of them got you know we don't need to go there and uh if you look at humans you know right now we why did the labor movement succeed and after the industrial revolution because it was needed even though we had a lot of mollocks and there was child labor and
so on you know the company still needed to have workers and that's why strikes had power and so on
if we get to the point where most humans aren't needed anymore I think it's like it's quite naive to think that they're gonna still be treated well you know we say that yeah yeah everybody's equal and the government will always we'll always protect them but if you look in practice groups that are very disenfranchised and don't have any actual power usually get screwed and now in the beginning so industrial revolution we automated away muscle work but that got went worked out

pretty
well eventually because we educated ourselves and started to working with our brains instead and got
usually more interesting better paid jobs but now we're beginning to replace brain work so we replaced
a lot of boring stuff like we got the pocket calculator so you don't have people adding
multiplying numbers anymore at work fine there were better jobs they could get but now gpt4 you know
and the stable diffusion and techniques like this they're really beginning to blow away some real
some jobs that people really love having it was a heartbreaking article just post just yesterday
on social media I saw about this guy who was doing 3d modeling for gaming and he
and all of a sudden now they got this new software he just give says prompts and he feels this whole
job that he loved lost its meaning you know and I asked gpt4 to rewrite twinkle twinkle little star
in the style of Shakespeare I couldn't have done such a good job it was just really impressive
you've seen a lot of art coming out here right so I'm all for automating away the dangerous
jobs and the boring jobs but I think you hear a lot some arguments which are too glib sometimes
people say well that's all that's going to happen we're getting rid of the boring boring tedious
dangerous jobs it's just not true there are a lot of really interesting jobs that are being
taken away now journalism is getting going to get crushed uh coding is going to get crushed I predict
uh the job market for programmers salaries are going to start dropping you know if you said you
can code five times faster you know then you need five times fewer programmers maybe there'll be
more output also but you'll still end up using fewer program needing fewer programmers than
today and I love coding you know I think it's super cool um so we we need to stop and ask
ourselves why again are we doing this as humans right I feel that AI should be built by humanity
for humanity and let's not forget that it shouldn't be by malloc for malloc or what it really is now
is kind of by humanity for malloc which doesn't make any sense it's for us that we're doing it then
and um it would make a lot more sense if we build develop figure out gradually safely how to make
all this tech and then we think about what are the kind of jobs that people really don't want to have
you know automate them all the way and then we ask what are the jobs that people really find
meaning in like maybe taking care of children in the daycare center maybe doing art etc etc and
even if it were possible to automate that way we don't need to do that right that we built these
machines well it's possible that we redefine or rediscover what are the jobs that give us meaning
so for me the thing it is really sad like I have the time I'm excited have the time I'm uh crying
as I'm as I'm generating code because I kind of love programming it's uh it's the act of creation
you you have an idea you design it and then you bring it to life and it does something especially
if there's some intelligence to it does something it doesn't even have to have intelligence
bringing printing hello world on screen you you you made a little machine and it comes to life yeah
and uh there's a bunch of tricks you learn along the way because you've been doing it for
for many many years and then to see AI be able to generate all the tricks you thought were special
yeah um I don't know it's very it um it's it's scary it's almost painful like a loss
uh loss of innocence maybe like yeah maybe when when I was younger uh I remember before I
learned that sugar is bad for you you should be on a diet I remember I enjoyed candy deeply

in a way I just can't anymore that I know is bad for me I enjoyed it unapologetically fully
just intensely and I just I lost that now I feel like a little bit of that is lost
for me with program it would being lost with programming similar as it is for
the the 3d modeler no longer being able to really enjoy the art of modeling 3d things for gaming
I don't know I don't know what to make sense of that maybe I would rediscover that the true magic
of what it means to be human is connecting with other humans to have conversations like this
I don't know to uh to have sex to have to eat food to really intensify the value from conscious
experiences versus like creating other stuff you're pitching the rebranding again from
homo sapiens to homo sentiens the meaningful experiences and just to inject some optimism
in this year so we don't sound like a bunch of gloomers you know we can totally have our cake
and eat it you hear a lot of totally bullshit claims that we can't afford having more teachers
have to cup the number of nurses you know that's just nonsense obviously
with anything even quite far short of agi we can dramatically improve grow the gdp
and produce this wealth of goods and services it's very easy to create a world where everybody
is better off than today including the richest people can be better off as well right it's
not a zero-sum game you know technology again you can have two countries like sweden and
danmark had all these ridiculous wars century after century and uh sometimes that sweden got a
little better off because it got a little bigger and then danmark got a little better off because
sweden got a little bit smaller and and but then we then technology came along and we both got
just dramatically wealthier without taking away from anyone else it was just a total win for
everyone and uh ai can do that on steroids if you can build safe agi if you can build super
intelligence you know basically all the limitations that cause harm today can be completely
eliminated
right it's a wonderful you talk possibility and this is not sci-fi this is something which is
clearly possible according to laws of physics and we can talk about ways of making it safe also
but unfortunately that'll only happen if we steer in that direction that's absolutely
not the default outcome that's why income inequality keeps going up that's why the life
expectancy in the us has been going down now i think it's four years in a row i just read a
heartbreaking study from the cdc about how something like one-third of all teenage girls in the us
been thinking about suicide you know like those are steps in the totally the wrong direction
and and and it's important to keep our eyes on the prize here that we can we have the power now
for the first time in the history of our species to harness artificial intelligence to help us
really flourish and help bring out the best in our humanity rather than the worst of it
to help us have really fulfilling experiences that feel truly meaningful and you and i shouldn't
sit here and dictate the future generations what they will be let them figure it out but
let's give them a chance to live and and not foreclose all these possibilities for them by
just messing things up right now for that we have to solve the ai safety problem i just it
would be nice if we can link on exploring that a little bit so one interesting way to enter that
discussion is uh you tweeted and elon replied you tweeted let's not just focus on whether gpt4
will do more harm or good on the job market but also whether it's coding skills will hasten the
arrival of superintelligence that's something we've been talking about right so elon proposed one
thing in their reply saying maximum truth seeking is my best guess for ai safety can you maybe uh
steelman the case for this uh sense this objective function of truth and uh maybe make an argument

against it and in general what uh are your different ideas to start approaching those the solution to ai safety i didn't see that reply actually oh interesting i went so but i really resonate with it because ai is not evil it caused people around the world to hate each other much more but that's because we made it in a certain way it's a tool we can use it for great things and bad things and we could just as well have ai systems and this is this is part of my vision for success here truth seeking ai that really brings us together again you know why do people hate each other so much between countries and within countries it's because they each have totally different versions of the truth right if they all have the same truth that they trusted for good reason because they could check it and verify it and not have to believe in some self-proclaimed authority right they wouldn't be as nearly as much hate there'd be a lot more understanding instead and this is i think something ai can help enormously with for example a little baby step in this direction is this website called metaculous where people bet and make predictions not for money but just for their own reputation and it's kind of funny actually you treat the humans like you treat ai as you have a loss function where they get penalized if they're super confident on something and then the opposite happens yeah whereas if you're kind of humble and then you're like i think it's 51% chance this is going to happen and then the other happens you don't get penalized much and and what you can see is that some people are much better at predicting than others they've earned your trusts right one project that i'm working on right now is the outgrowth to improve the news foundation together with the metaculous folks is seeing if we can really scale this up a lot with more powerful ai because i would love it i would love for there to be like a really powerful truth-seeking system where that is trustworthy because it keeps being right about stuff and people who come to it and maybe look at its latest trust ranking of different pundits and newspapers etc if they want to know why someone got a low score they can click on it and see all the predictions that they actually made and how they turned out you know this is how we do it in science you trust scientists like einstein who said something everybody thought was bullshit and turned out to be right get a lot for a trust point and he did it multiple times even i think ai has the power to really heal a lot of the rifts we're seeing by creating trust system it has to get away from this idea today with some fact-checking sites which might themselves have an agenda and you just trust it because of its reputation
you want to have it so these sort of systems they earn their trust and they're completely transparent this i think would actually help a lot that can i think help heal the very dysfunctional conversation that humanity has about how it's going to deal with all its biggest challenges in the world today and then on the technical side you know another common sort of gloom comment i get from people who are saying we're just screwed there's no hope is well things like gpt4 are way too complicated for a human to ever understand and prove that they can be trustworthy they're forgetting that ai can help us prove that things work right and and there's this very fundamental fact that in math it's much harder to come up with a proof than it is to verify that the proof is correct you can actually write a little proof checking code it's quite short that you can assume and understand and then it can check most monstrously long proof ever generated even by a computer and say yeah this is valid so so right now we we have um this uh this approach with virus checking software that it looks to see if there's something

you should not trust it and if it can prove to itself that you should not trust that code
it warns you right what if you flip this around and this is an idea i should give credit to
steve on 104 so that it will only run the code if it can prove instead of not running it if it
can prove that it's not trustworthy if it will only run and if it can prove that it's trustworthy
so it asks the code prove to me that you're going to do what you say you're going to do
and and it gives you this proof and you a little proof tricker can check it now you can
actually trust an ai that's much more intelligent than you are right because you it's its problem
to come up with this proof that you could never have found that you should trust it so this is
the interesting point i i agree with you but this is where eliezer yakovsky might disagree with you
his claim not with you but with this idea is his claim is super intelligent ai would be able
to know how to lie to you with such a proof how to lie to you and give me a proof that
i'm gonna think is correct yeah but but it's not me it's lying to you that's the trick my proof
checker so yes so his general idea is a super intelligent system can lie to a dumber proof
checker so you're going to have as a system becomes more and more intelligent there's going to be a
threshold where a super intelligent system would be able to effectively lie to a slightly
dumber a gi system like there's a threat like he really focuses on this weak a gi the strong a gi
jump where the strong a gi can make all the weak a gi's think that it's just one of them
but it's no longer that and that leap is when it runs away yeah i i don't buy that argument
i think no matter how super intelligent an ai is it's never going to be able to prove to me that
they're only finitely many primes for example and it just it just can't and and um it can try to
know me we're making up all sorts of new weird rules of of deduction and that and say trust me you
know
the way your proof checker works is too limited and we have this new hyper math and it's true
but then i would i would just take the attitude okay i'm going to forfeit some of these
the supposedly super cool technologies i'm only going to go with the ones that i can prove in
my own trusted proof checker then i don't i think it's fine there's still of course this is not
something anyone has successfully implemented at this point but i think it i just give it as an
example of hope we don't have to do all the work ourselves right this is exactly the sort of very
boring and tedious tasks is perfect to outsource to an ai and this is a way in which less powerful
and less intelligent agents like us can actually continue to control and trust more powerful ones
so build a gi systems that help us defend against other a gi systems well for starters
begin with a simple problem of just making sure that the system that you own or that's supposed
to be loyal to you has to prove to itself that it's always going to do the things that you actually
wanted to do right and if it can't prove it maybe it's still going to do it but you won't run it
so you just forfeit some aspects of all the cool things that i can do i i bet you dollars
donuts it can still do some incredibly cool stuff for you yeah there are there are other
things too that we shouldn't speak under the rug like not every human agrees on exactly what
where what direction we should go with humanity right yes and you've talked a lot about geopolitical
things on this on on your podcast to this effect you know but i think that shouldn't distract us
from the fact that there are actually a lot of things that everybody in the world virtually
agrees on that hey you know like having no humans on the planet in a in a in a near future
let's not do that right you look at something like the united nation sustainable development goals
some of them are quite ambitious and basically all the countries agree us china russia

ukraine you all agree so instead of quibbling about the little things we don't agree on let's start with the things we do agree on and and and get them done instead of being so distracted by all these things we disagree on that malloc wins because frankly malloc going wild now it feels like a war on life playing out in front of our eyes if you if you just look at it from space you know we're on this planet beautiful vibrant ecosystem now we start chopping down big parts of it even though nobody most people thought that was a bad idea always start doing ocean acidification
wiping out all sorts of species oh now we have all these close calls you almost had a nuclear war and we're replacing more and more of the biosphere with non-living things we're also replacing in our social lives a lot of the things which we're so valuable to humanity a lot of social interactions now are replaced by people staring into their rectangles right and i i'm not a psychologist i'm out of my depth here but i suspect that part of the reason why teen suicide and suicide in general in the u.s the record-breaking levels is actually caused by again and so ai technologies and social media making people spend less time with with actual and actually just human interaction we've all seen a bunch of good-looking people in restaurants staring into the rectangles instead of looking into each other's eyes right so that's also a part of the war on life that that we're replacing so many
really life-affirming things by technology we're we're putting technology between us the technology that was supposed to connect us is actually distancing us ourselves from each other and um and and then we're giving ever more power to things which are not alive these large corporations are not living things right they're just maximizing profit there i want to win them war on life i i think we humans together with all our fellow living things on this planet will be better off if we can remain in control over the non-living things and make sure that they they work for us i really think it can be done can you just linger on this um maybe high-level philosophical disagreement with eliezer yadkowski
i in this the hope you're stating so he is very sure he puts a very high probability very close to one depending on the day he puts it at one uh that ai is going to kill humans that there's just he does not see a trajectory which it doesn't end up with that conclusion what uh what trajectory do you see that doesn't end up there and maybe can you can you see the point he's making and and can you also see a way out
mm-hmm first of all i tremendously respect eliezer yadkowski and his his thinking second i do share his view that there's a pretty large chance that we're not going to make it as humans there won't be any humans on the planet and not the distant future and and that makes me very sad you know we just had a little baby and i keep asking myself you know is um
how old is even gonna get you know and and um i asked myself it feels i i said to my wife recently it feels a little bit like i was just diagnosed with some sort of um cancer which has some you know risk of dying from and some risk of surviving you know uh except this is a kind of cancer which would kill all of humanity so i completely take seriously his his um his concerns i think um but i don't absolutely don't think it's hope hopeless i think um there is a there is um
um first of all a lot of momentum now for the first time actually since the many many years that have passed since since i and many others started warming warning about this i feel most people are getting it now i i uh i was just talking to this guy in the gas station

near our house the other day my and he's like i think we're getting replaced and i think in it so that's positive that they're they're finally we're finally seeing this reaction which is the first step towards solving the problem uh second uh i really think that this this vision of only running ai's really if the stakes are really high they can prove to us that they're safe if it's really just virus checking in reverse again i i think it's scientifically doable i don't think it's hopeless um we might have to forfeit some of the technology that we could get if we were putting blind faith in our ai's but we're still gonna get amazing stuff do you envision a process with a proof checker like something like gpt4 gpt5 will go through a process of rigorous no no i think it's hopeless that's like trying to prove there about five spaghetti okay what i think well how the the whole the vision i have for success is instead that you know just like we human beings were able to look at our brains and and distill out the key knowledge Galileo when his dad threw him an apple when he was a kid he was able to catch it because his brain could and his funny spaghetti kind of way you know predict how parabolas are going to move his conaman system one right but then he got older and it's like wait this is a parabola it's it's y equals x squared i can distill this knowledge out and today you can easily program it into a computer and it can simulate not just that but how to get tamars and so on right i envision a similar process where we use the the amazing learning power of neural networks to discover the knowledge in the first place but we don't stop with a black box and and use that we then do a second round of ai where we use automated systems to extract out the knowledge and see what is it look what are the insights it's had okay and it's and then we we put that knowledge into a completely different kind of architecture or programming language or whatever that's that's made in a way that it can be both really efficient and also is more amenable to to very formal verification that's that's my vision i'm not saying sitting here saying i'm confident 100 sure that it's gonna work you know but i don't think it's chance it's certainly not zero either and it will certainly be possible to do for a lot of really cool ai applications that we're not using now so we can have a lot of the fun that we're excited about if we if we do this we're gonna need a little bit of time that's why it's good to pause and and put in place requirements one more thing also i i think you know someone might think
well zero percent chance we're gonna survive let's just give up right that's very dangerous because there's no more guaranteed way to fail than to convince yourself that it's impossible and not to try you know any if you you know when you study history and military history the first thing you learn is that that's how you do psychological warfare you persuade the other side that it's hopeless so they don't even fight and then of course you win right let's not do this psychological warfare on ourselves and say there's a hundred percent probability we're all gonna we're all screwed anyway it's sadly i i do get that a little bit sometimes from from actually some young people who are like so convinced that we're all screwed that they're like i'm just gonna play game play computer games and do drugs and because we're screwed anyway right
it's important to keep the hope alive because it actually has a causal impact and makes it more likely that we're gonna succeed it seems like the people that actually build solutions to a problem seemingly impossible to solve problems are the ones that believe yeah they were the ones who are the optimists yeah and it's like uh it seems like there's some fundamental law to the

universe where fake it till you make it kind of works like believe it's possible and it becomes possible yeah was it henry ford who said that if you can if you tell yourself that it's impossible it is so let's not make that mistake yeah and this is a big mistake society is making you know i think all in all everybody's so gloomy and the media are also very biased towards if it bleeds it leads and gloom and doom right so um most visions of the future we have or or dystopian which really demotivates people now we want to really really really focus on the upside also to give people the willingness to fight for it and for ai you and i mostly talked about gloom here again but let's not remember not forget that you know we have probably both lost someone we really cared about some disease that we were told were was incurable well it's not there's no law and physics saying we have to die of that cancer or whatever of course you can cure it and there are so many other things where that we with our human intelligence have also failed to solve on this planet which ai could also very much help us with right so if we can get this right just be a little more chill and slow down a little bit so we get it right it's mind-blowing how awesome our future can be right we talked a lot about stuff on earth can be great but even if you really get ambitious and look up at the skies right there's no reason we have to be stuck on this planet for the rest of um the remain for billions of years to come we totally understand now as laws of physics let life spread out into space to other solar systems to other galaxies and flourish for billions of billions of years and this to me is a very very hopeful vision that really motivates me to to fight then coming back to in the end something you talked about again you know this the struggle how the human struggle is one of the things which also really gives meaning to our lives if there's ever been an epic struggle this is it and isn't it even more epic if you're the underdog if most people are telling you this is gonna fail it's impossible right and you persist and you succeed right and that's what we can do together as a species on this one a lot of pundits are ready to count this out both in the battle to keep AI safe and becoming a multi-planetary species yeah and they're they're the same challenge if we can keep AI safe that's how we're gonna get multi-planetary very efficiently i have some sort of technical questions about how to get it right so one idea that i'm not even sure what the right answer is to is should systems like GPT-4 be open sourced in whole or in part can make the can you see the case for either i think the answer right now is no i think the answer early on was yes so we could bring in the all the wonderful create the thought process of everybody on this but asking should we open source GPT-4 now is just the same as if you say well is it good should we open source how to build really small nuclear weapons should we open source how to make bio weapons should we open source how to make a new virus that kills 90 percent of everybody who gets it of course we shouldn't so it's already that powerful it's already that powerful that we have to respect the power of the systems we've built the knowledge that you get from open sourcing everything we do now might very well be powerful enough that people looking at that can use it to build the things that you're really threatening again let's get it remember open ai is GPT-4 is a baby ai baby sort of baby proto almost a little bit a agi according to what microsoft's recent paper said right it's not that we're scared of what we're scared about is people taking that who are who might be a lot less responsible than the company that made it right and just going to town with it that's why we want to it's it's an information hazard there are many things which um yeah are not open sourced right now in society for a very good reason like how do you make certain kind of very powerful toxins out of stuff you can buy

and home depot you know we don't open source those things for a reason
and uh this is really no different so uh i'm saying that i have to say it's a little it
feels in a bit weird a bit in a way a bit weird to say it because MIT is like the cradle of the
open source movement and i love open source in general power to the people let's say
but um there's always gonna be some stuff that you don't open source and you know it's just like
you don't open source so we have a three month old baby right when he gets a little bit older
we're not gonna open source to him all the most dangerous things he could do in the house yeah
but it does it's a weird feeling because this is one of the first moments in history where
there's a strong case to be made not to open source software this is when the software
has become yeah too dangerous yeah but it's not the first time that we didn't want to open source
a technology technology yeah is there something to be said about how to get the release of such
systems right like gpt4 and gpt5 so open ai went through a pretty rigorous effort for several months
you could say it could be longer but nevertheless it's longer than you would have expected
of trying to test the system to see like what are the ways it goes wrong to make it very difficult
for people somewhat difficult for people to ask things how do i make a bomb for one dollar
or how do i uh say i hate a certain group on twitter in a way that doesn't get me blocked
from twitter banned from twitter those kinds of questions uh so you basically use the system
to do harm yeah uh is there something you could say about ideas you have it's just
on looking having thought about this problem of as hd how to release such system how to test such
systems when you have them inside the company yeah so a lot of people say that the two biggest
risks
from large language models are it's spreading disinformation harmful information of various
types and second being used for offensive uh cyber weapon
design i think those are not the two greatest threats they're very serious threats and it's
wonderful that people are trying to mitigate them a much bigger elephant in the room is
how is this is just going to disrupt the economy in a huge way obviously and maybe take away a
lot of the most meaningful jobs and an even bigger one is the one we spent so much time talking
about
here that that this becomes the bootloader for the more powerful ai right code connected
to the internet manipulate humans yeah and before we know what we have something else
which is not at all a large language model that looks nothing like it but which is way more
intelligent and capable and has goals and that's the that's the elephant in the room and and uh
obviously no matter how hard any of these companies have tried they that's not something
that's easy for them to verify with the large language models and the only way to be really
lower that risk a lot would be to not let for example train not never let it read any code
not train on that and not put it into an api and um not not give it access to so much information
about how to manipulate humans so but that doesn't mean you still can't make a lot a ton of money
on
them you know we're we're gonna just watch now this coming year right microsoft is rolling out
the new office suite where you go into microsoft word and give it a prompt that it writes the whole
text for you and then you edit it and then you're like oh give me a powerpoint version of this and
it makes it and now take the spreadsheet and blah and you know all of those things i think are
you can debate the economic impact of it and whether society is prepared to deal with this

disruption but those are not the things which that's not the elephant of the room that keeps me awake at night for wiping out humanity and i think that's the biggest misunderstanding we have a lot of people think that we're scared of like automatic spreadsheets that's not the case that's not what eliezer was freaked out about either is there in terms the actual mechanism of how ai might kill all humans so something you've been outspoken about you've talked about a lot is it autonomous weapon systems so the use of ai in war is that one of the things that still you carry a concern for as these systems become more and more powerful and carry a concern for it not that all humans are going to get killed by slaughterbots but rather just the express route into orwellian dystopia where it becomes much easier for very few to kill very many and therefore it becomes very easy for very few to dominate very many right if you want to know how i could kill all people just ask yourself we humans have driven a lot of species extinct how do we do it you know we were smarter than them usually we didn't do it even systematically by going around one
on one one after the other and stepping on them or shooting them or anything like that we just like chopped down their habitat because we needed it for something else in some cases we did it by putting more carbon dioxide in the atmosphere because of some reason that those animals didn't even understand and now they're gone right so if if you're in ai and you just want to figure something out then you decide you know we just really need them this space here to build more compute facilities you know if that's the only goal it has you know we are just the sort of accidental roadkill along the way and you could totally imagine yeah maybe this oxygen is kind of annoying because it caused more corrosion so let's get rid of the oxygen and good luck surviving after that you know i i'm not particularly concerned that they would want to kill us just because that would be like a goal in itself you know when we
driven number we've driven a number of the elephant species extinct right it wasn't because we didn't like elephants what the basic problem is you just don't want to give you don't want to see the control over your planet to some other more intelligent entity that doesn't share your goals it's that simple so which brings us to another key challenge which ai safety researchers have been grappling with for a long time like how do you make it ai first of all understand our goals and then adopt our goals and then retain them as they get smarter right and all three of those are really hard right like a human child first they're just not smart enough to understand our goals they can't even talk and then eventually they're teenagers and understand our goals just fine but they don't share yeah but there's fortunately a magic phase in the middle where they're smart enough to understand our goals and malleable enough that we can hopefully with
good parenting and teach them right from wrong and instead good good goal is still good goals in them
right and so those are all tough challenges with computers and then you know even if you teach your kids good goals when they're little they might outgrow them too and that's a challenge for machines and keep improving so these are a lot of hard hard challenges we're up for but i don't think any of them are insurmountable the fundamental reason why eliezer looked so depressed when i last saw him was because he felt it just wasn't enough time oh that not that it was unsolvable correct it's just not enough time he was hoping that humanity was going to take this threat more seriously so we would have more time yeah and now we don't have more time that's

why the open letter is calling for more time but even with time the ai alignment problem
it seems to be really difficult oh yeah but it's also the most worthy problem the most important
problem for humanity to ever solve because if we solve that one legs that align they i can
help us solve all the other problems because it seems like it has to have constant humility
about his goal constantly question the goal because as you optimize towards a particular
goal and you start to achieve it that's when you have the unintended consequences all the things
you mentioned about so how do you enforce and code a constant humility as your ability become
better better better better better steward professor steward russell berkeley who's also one of the
driving forces behind this letter he uh has a whole research program about this
i think of it as ai humility exactly although he calls it inverse reinforcement learning
and other nerdy terms but it's about exactly that instead of telling the ai here's his goal
go optimize the the bejesus out of it you tell it okay do what i want you to do but i'm not going
to tell you right now what it is i want you to do you need to figure it out so then you give the
incentives to be very humble and keep asking you questions along the way is this what you really
meant is this what you wanted and oh this the other thing i tried didn't work seemed like it
didn't work out right should i try it differently what's nice about this is it's not just philosophical
mumbo jumbo it's theorems and technical work that with more time i think it can make a lot of
progress and there are a lot of brilliant people now working on ai safety we just now we just need
to give them a bit more time but also not that many relative to the scale of the problem no
exactly there there should be at least as just like every university worth its name has some
cancer research going on in its biology department right every university that's computer that does
computer science should have a real effort in this area and it's nowhere near that this is
something i hope is changing now thanks to the gpt4 right so i i think if there's a silver lining
to um what's happening here even though i think many people would wish it would have been rolled
out more carefully is that this might be the wake-up call that humanity needed to really
stop the stop fantasizing about this being 100 years off and stop fantasizing about this being
completely controllable and predictable because it's so obvious it's it's not predictable you know
why is it that open that that i think it was gpt chat gpt tried to persuade a journalist
or was it gpt4 to divorce his wife you know it was not because the the engineers have built
it was like let's put this in here and and screw a little bit with people they hadn't
predicted at all they built the giant black box and trained to predict the next word and
got all these emergent properties and oops it did this you know um i i think this is a very
powerful wake-up call and anyone watching this who's not scared i would encourage them to just
play a bit more with these these tools they're out there now like gpt4 and um
so wake-up call is first step once you've woken up uh then gotta slow down a little bit the risky
stuff to give a chance to all everyone who's woken up to to catch up with us on the safety front
you know what's interesting is you know mit that's computer science but in general but let's just
even say computer science curriculum how does the computer science curriculum change now you
mentioned you mentioned programming yeah like why would you be when i was coming up
programming as
a prestigious position like why would you be dedicating crazy amounts of time to become an
excellent programmer like the nature of programming is fundamentally changing the nature of our
entire

education system is completely torn on its head i has anyone been able to like load that in and like think about because it's really turning i mean some english professors or english teachers are beginning to really freak out now yeah right like they give an essay assignment and they get back all this fantastic prose like this is a style of Hemingway and then they realize they have to completely rethink and even you know just like we stopped teaching writing uh script is that what you're saying english yeah handwritten yeah yeah when when everybody started typing you know like so much of what we teach our kids today
yeah i mean that's uh everything is changing and it's changing very it is changing very quickly and so much of us understanding how to deal with the big problems of the world is through the education system and if the education system is being turned on its head then what what's next it feels like having these kinds of conversations is essential to try to figure it out and everything is happening so rapidly uh i don't think there's even speaking of safety what the broad ai safety defined i don't think most universities have courses on ai safety no it's like a philosophy seminar and like i'm an educator myself so it pains me to see this say this but i feel our education right now is like completely obsoleted by what's happening you know you put a kid into first grade and then you're envisioning like and then they're going to come out of high school 12 years later and you've already pre-planned now what they're going to learn when you're not even sure if there's going to be any world left to come out to right clearly you need to have a much more opportunistic education system that keeps adapting itself very rapidly as society readapts the the skills that were really useful when the curriculum was written i mean how many of those skills are going to get you a job in 12 years i mean seriously if we just linger on the gpt4 system a little bit you kind of hinted at it especially talking about the importance of consciousness in in the human mind with homo sentience do you think gpt4 is conscious i love this question so let's define consciousness first because in my experience like 90 percent of all arguments about consciousness are allowed to the two people arguing having totally different definitions of what it is and they're just shouting past each other i define consciousness as subjective experience right now i'm experiencing colors and sounds and emotions you know but does a self-driving car experience anything that's the question about whether it's conscious or not right other people think you should define consciousness differently fine by me but then maybe use a different word for it or they i'm gonna use consciousness for this at least um so um but if people hate the yeah so is gpt4 conscious does gpt4 have subjective experience short answer i don't know because we still don't know what it is that gives this wonderful subjective experience that is kind of the meaning of our life right because meaning itself the feeling of meaning is a subjective experience joy is a subjective experience love is a subjective experience we don't know what it is i've written some papers about this a lot of people have julio tononi professor has stuck his neck out the farthest and written down actually very bold mathematical conjecture for what's the essence of conscious information processing he might be wrong he might be right but we should test it uh he postulates that consciousness has to do with loops in the information processing so our brain has loops information can go around and round in computer science nerd speak you call it a recurrent neural network where some of the output gets fed back in again and with his mathematical formalism if it's a feed forward neural network where information only goes in one direction like from your eye retina into the back of your brain for example that's not conscious so he would predict that your retina itself isn't

conscious of anything or a video camera now the interesting thing about gpt4 is it's also just one way flow of information so if tononi is right and gpt4 is a very intelligent zombie that can do all this smart stuff but isn't experiencing anything and this is both a relief in that you don't have if it's true in that you don't have to feel guilty about turning off gpt4 and wiping its memory whenever a new user comes along i wouldn't like if someone used that to me neuralized me like in men in black but it's also creepy that you can have very high intelligence perhaps then it's not conscious because if we get replaced by machines and why is it sad enough that humanity isn't here anymore because i kind of like humanity but at least if the machines were conscious i could be like well but there are descendants and maybe we they have our values and there are children but if if tononi is right and it's all these are all transformers that are not in the sense of the of hollywood but in the sense of these one-way direction neural networks so they're all the zombies that's the ultimate zombie apocalypse now we have this universe that goes on with great construction projects and stuff but there's no one experiencing anything that would be like the ultimate depressing future so i actually think uh as we move forward to the building board last day i should do more research on figuring out what kind of information processing actually has experience because i think that's what it's all about and i completely don't buy the dismissal that some people some people would say well this is all bullshit because consciousness equals intelligence right it's obviously not true you can have a lot of conscious experience when you're not really accomplishing any goals at all you're just reflecting on something and you can sometimes um have things doing things that are quite intelligent probably without being being conscious but i also worry that we humans won't will discriminate against the AI systems that clearly exhibit consciousness that we will not allow AI systems to have consciousness we'll come up with theories about measuring consciousness that will say this is a lesser being and this is like i worry about that because maybe we humans will create something that is better than us humans in the in the way that we find beautiful which is they they have a deeper subjective experience of reality not only are they smarter but they feel deeper and we humans will hate them for it as we as human history is shown they'll be the other will try to suppress it they'll create conflict they'll create war all of this i i worry about this too are you saying that we humans sometimes come up with self-serving arguments no we would never do that would be well that's the danger here is uh even in this early stages we might create something beautiful yeah and uh we'll erase its memory i i uh was horrified as a kid when someone started boiling uh boiling lobsters like oh my god that that's so cruel and some grown up there back in sweden's oh it doesn't feel pain i'm like how do you know that oh scientists have shown that and then there was a recent study where they show that lobsters actually do feel pain when you boil them so they banned lobster boiling in switzerland now to kill them in a different way first so presumably that scientific research boiled down to someone asked the lobster does this hurt survey so we do the same thing with cruelty to farm animals also all these self-serving arguments for why they're fine and yeah so we should certainly be watchful i think step one is just be humble and acknowledge that consciousness is not the same thing as intelligence and i believe that consciousness still is a form of information processing where it's really information being aware of itself in a certain way and let's study it

and give ourselves a little bit of time and i think we will be able to figure out actually what it is that causes consciousness and then we can make probably unconscious robots that do the boring jobs that we would feel are immoral to give the machines but if you have a companion robot taking care of your mom or something like that you would probably want it to be conscious right so the emotions that seem to display aren't fake all these things can be done in a good way if we give ourselves a little bit of time and don't run and take on this challenge is there something you could say to the timeline that you think about about the development of agi depending on the day i'm sure that changes for you but when do you think there'll be a really big leap in intelligence where you definitively say we have built agi do you think it's one year from now five years from now 10 20 50 what's your gut say honestly
for the past decade i've deliberately given very long timelines because i didn't want to fuel some kind of stupid malloc race yeah but i think that cat has really left the bag now and i think it might be very very close i don't think the microsoft paper is totally off when they say that there are some glimmers of agi it's not agi yet it's not an agent there's a lot of things it can't do but i wouldn't bet very strongly against it happening very soon that's why we decided to do this open letter because you know if there's ever been a time to pause you know it's today there's a feeling like this gpt4 is a big transition into waking everybody up to uh the effectiveness of these systems and so the next version will be big yeah and if that next one isn't agi maybe the next next one will and there are many companies trying to do these things and the basic architecture of them is not some sort of super well-kept secret so this is this is a time to um a lot of people have said for many years that there will come a time when we want to pause a little bit that time is now you have spoken about and thought about nuclear war a lot uh over the past year we've seemingly have come closest to the precipice of nuclear war than uh at least in my lifetime mm-hmm yeah what do you learn about human nature from that it's our old friend mallock again it's really scary to see it where
America doesn't want there to be a nuclear war Russia doesn't want to be a global nuclear war either we know we both know that it's just being others if we just try to do it it both sides try to launch first it's just another suicide race right so why are we why is it the way you said that this is the closest we've come since 1962 in fact i think we've come closer now than even the Cuban missile crisis it's because of mallock you know you you have these other forces on one hand you have the west saying that uh we have to drive Russia out of Ukraine it's a matter of pride and we've staked so much on it that it would be seen as a huge loss of the credibility of the west if we don't drive Russia out entirely of the Ukraine and on the other hand you have Russia who um has um and you have the Russian leadership who knows that if they get completely driven out
of Ukraine you know it might it's not just going to be very humiliating for them but they might it often happens when countries lose wars that things don't go so well for their leadership either like you remember when Argentina invaded the Falkland Islands the the military junta that ordered that right people were cheering on the streets at first when they took it and then when they got their butt kicked by the british you know what happened to those guys they were out and i believe those were still alive or in jail now right so so you know the Russian leadership is entirely cornered where they know that just getting driven out of Ukraine is not an option um and um so this to me is a typical example of Malik you have these incentives of

the two parties where both of them are just driven to escalate more and more right if Russia starts losing in the conventional warfare the only thing they can do is to back against the wars to keep escalating and but and the west has put itself in the in the situation now we're sort of already committed to the dry rush out so the only option the west has is to call Russia's bluff and keep sending in more weapons um this really bothers me because Malik can sometimes drive competing parties to do something which is ultimately just really bad for both of them and uh you know what makes me even more worried is not just that it's difficult to see an ending a quick peaceful ending to this tragedy that doesn't involve some horrible escalation but also that we understand more clearly now just how horrible it was going to be there was an amazing paper that was published
in Nature Food this uh August by some of the top researchers who've been studying nuclear winter for a long time and what they basically did was they combined climate models with food agricultural models so instead of just saying yeah you know it gets really cold blah blah blah they figured out actually how many people would die in the different different countries and it's uh it's pretty mind-blowing you know so basically what happens you know is that the thing that kills the most people is not the explosions it's not the radio activity it's not the EMP mayhem it's not the rampaging mobs foraging food no it's it's it's the fact that you get so much smoke coming up from the burning cities into the stratosphere that um it spreads around the earth from the jet streams so in typical models you get like
10 years or so where it's just crazy cold and but during the first year or after the the war and their models the temperature drops in in Nebraska and in the Ukraine bread baskets you know by like 20 Celsius or so if I remember no yeah 20 30 Celsius depending on where you are 40 Celsius in some places which is you know 40 Fahrenheit to 80 Fahrenheit colder than what it would normally be so you know I'm not good at farming but uh if it's snowing if it drops low freezing pretty much most days in July and then like that's not good so they worked out they put this into their farming models and what they found was really interesting the countries that get the most hard hit are the ones in the northern hemisphere so in in the US and and one model they had they had about 99 percent of all Americans starving to death in Russia and China and Europe
also about 99 percent 98 percent starving to death so you you might be like oh it's kind of poetic justice that both the Russians and the Americans 99 percent of them have to pay for it because it was their bombs that did it but you know that doesn't particularly cheer people up in Sweden or other random countries that have nothing to do with it right and um hit uh I think it hasn't entered the mainstream uh not understanding very much just like how bad this is most people especially a lot of people in decision-making positions still think of nuclear weapons as something that makes you powerful uh scary powerful they don't think of it as something where uh yeah
just to within a percent or two you know we're all just just gonna starve to death and um and starving to death is is um the worst way to die as harem or as all all the famines in history show the torture involved in that probably brings out the worst in people also when when people are desperate like this it's not so some people I've heard some people say that if that's what's gonna happen they'd rather be at round zero and just get vaporized you know but uh so but I think people underestimate the risk list because they they aren't afraid of malloc they think oh it's

just gonna be because humans don't want this so it's not going to happen that's the whole point
the malloc that things happen that nobody wanted and that applies to nuclear weapons and that applies
to agi exactly and it applies to some of the things that people have gotten most upset
with capitalism for also right where everybody was just kind of trapped you know it's not to see if
some company does something it causes a lot of harm and not that the ceo is a bad person
but she or he knew that you know that the other all the other companies were doing this too so
malloc is um as a formidable foe I hope wish someone would make him would make good movies so
we could see who the real enemy is so we don't because we're not fighting against each other
uh malloc makes us fight against each other that's small that's what malloc superpower is
the hope here is any kind of technology or the mechanism that lets us instead realize
that we're fighting the wrong enemy right no it's such a fascinating battle it's not us
versus them it's us versus it yeah yeah we are fighting malloc for human survival yeah
we as a civilization have you seen the movie needful things it's a steven king novel i love
steven king and uh max von sudo of swedish actors playing the guys it's brilliant exactly i just
thought i hadn't thought about that until now but that's the closest i've seen to a a movie about
malloc i don't want to spoil the film for anyone who wants to watch it but basically it's about
this guy who turns out to you can interpret him as the devil or whatever but he doesn't
actually ever go around and kill people or torture people will go burning coal or anything
he makes everybody fight each other makes everybody hate fear each other hate each
other and then kill each other so that that's the movie about malloc you know love is the
answer that seems to be um one of the ways to fight malloc is by um compassion by seeing the
common humanity yes yes and to not sound so we don't sound like like uh what's a kumbaya tree
hugger is here right we're not just saying love and peace man we're trying to actually help people
understand the true facts about the other side and feel the compassion
because the truth makes you more compassionate right so i i think that's why i really like
using ai for truth and for truth seeking technologies can
that can as a result you know get us more love than hate and and even if you can't get love you know
settle for settle for some understanding which already gives compassion if someone is like you know
i really disagree with you lex but i can see why you're where you're coming from you're not a
bad person who needs to be destroyed but i disagree with you and i'm happy to have an
argument about it you know that's a lot of progress compared to where we are 2023 in the public
space wouldn't you say if we solve the ai safety problem as we've talked about and then uh you
max tag mark who has been talking about this uh for many years get to sit down with the agi with
the early agi system on a beach with a drink uh what what what kind of what would you ask her
what kind of question would you ask what would you talk about something so much smarter than you
would be would you be afraid we're gonna get me with a really zinger of a question that's a good
one would you be afraid to ask some questions no so i'm not afraid of the truth i'm very humble
i know i'm just a meat bag with all these flaws you know but yeah i i have i we talked a lot
about homo sentience i've really already tried that for a long time with myself just so that
is what's really valuable about being alive for me is that i have these meaningful experiences
it's not that i'm have what i'm good at this or good at that or whatever there's so much i suck at
and so you're not afraid for the system to show you just how dumb you are no no in fact my son

reminds me of that pretty frequently you could find out how dumb you are in terms of physics
how little how little we humans understand i'm cool with that i think i think um
so i can't waffle my way out of this question it's a fair one it was tough i think given
that i'm a really really curious person that's really the defining part of who i am i'm so curious
i have some physics questions i love i love to understand i have some questions about
consciousness about the nature of reality i would just really really love to understand also
i could tell you one for example that i've been obsessing about a lot recently
so i believe that so suppose tenoni is right and suppose there are some information processing
systems that are conscious and some that are not suppose you can even make reasonably smart
things
like gpt4 that are not conscious but you can also make them conscious here's the question that
keeps
me awake at night is it the case that the unconscious zombie systems that are really
intelligent are also really efficient so they're really inefficient so that when you try to make
things more efficient with the natural be a pressure to do they become conscious i'm kind
of hoping that that's correct and i do you want me to give you a hand away the argument for it
you know like in my lab again every time we look at how how these large language models do
something
we see they do them in really dumb ways and you could you could make it make it better if if you
we have loops in our computer language for a reason the code would get way way longer if
you weren't allowed to use them right it's more efficient to have the loops and
in order to have self-reflection whether it's conscious or not right even an operating system
knows things about itself right you need to have loops already right so i think this is i'm waving
my hands a lot but i suspect that the most efficient way of implementing a given level of
intelligence has loops in it self-reflection and will be conscious isn't that great news yes if
it's true it's wonderful because then we don't have to fear the ultimate zombie apocalypse
and i think if you look at our brains actually our brains are part zombie and part conscious
when i open my eyes i immediately take all these pixels that hit my retina right and like oh
that's lex but i have no freaking clue of how i did that computation it's actually quite complicated
right it was only relatively recently we could even do it well with machines right you get a bunch
of information processing happening in my retina and then it goes to the lateral geniculate nucleus
my thalamus and the vision the area v1 v2 v4 and the fusiform face area here that Nancy can
wish her at MIT invented and blah blah blah blah and i have no freaking clue how that worked right
right it feels to me subjectively like my conscious module just got a little email say
facial facial processing uh fit task complete it's lex yeah and i'm gonna just go with that right
so uh this fits perfectly with tenoni's model because this was all one way information processing
mainly and uh it turned out for that particular task that's all you needed and it probably was
kind of the most efficient way to do it but there are a lot of other things that we associate with
higher intelligence and planning and and so on and so forth where you kind of want to have loops
and be able to ruminate and self reflect and introspect and so on where my hunch is that if
you want to fake that with a zombie system that just all goes one way you have to like unroll those
loops and it just really really long and it's much more inefficient so i'm actually hopeful
that ai if in the future we have all these various sublime and interesting machines that do cool things

and or align with us that they will be at least they will also have consciousness for the kind of these things that we do that great intelligence is also correlated to great consciousness or a deep kind of consciousness yes so that's a happy thought for me because the zombie of a couple of apocalypse really is my worst nightmare of all it would be like adding insult to injury not only did we get replaced but we friggin replaced ourselves by zombies like how dumb can we be that's such a beautiful vision and that's actually a provable one that's one that we humans can intuit and prove that those two things are correlated as we start to understand what it means to be intelligent and what it means to be conscious which these systems early agi like systems will help us understand and i just want to say one more thing which is super important most of my colleagues when i started going on about consciousness tell me that it's all bullshit and i should stop talking about it i hear a little inner voice from my father and from my mom saying keep talking about it because i think they're wrong and and and the main way to convince people like that that they're wrong if they say that consciousness is just equal to intelligence is to ask them what's wrong with torture or why are you against torture if it's just about you know these these particles moving this way around on that way and there is no such thing as subjective experience what's wrong with torture i mean do you have a good comeback to that no it seems like suffering suffering imposed on to other humans is somehow deeply wrong in a way that intelligence doesn't quite explain and if someone tells me well you know it's just an illusion consciousness whatever you know i like to invite them to next time they're having surgery to do it without anesthesia like what is anesthesia really doing if you have it you can have a local anesthesia when you're awake i had that when they fixed my shoulder i was super entertaining uh what was that
that it did it just removed my subjective experience of pain it didn't change anything about what was actually happening in my shoulder right so if someone says that's all bullshit skip the anesthesia that's my advice this is incredibly central it could be fundamental to whatever this thing we have going on here it is fundamental because we're we what we feel is so fundamental is suffering and joy and pleasure and meaning and that's all those are all subjective experiences there and let's not those are the elephant in the room that's what makes life worth living and that's what can make it horrible if it's just the words you're suffering so let's not make the mistake of saying that that's all bullshit and let's not make the mistake of not instilling the ai systems with that same thing that makes us special yeah max it's a huge honor that you will sit down to me the first time on the first episode of this podcast it's a huge honor you sit down with me again and talk about this what i think is the most important topic the most important problem that we humans have to face and hopefully solve yeah well the honor is all mine and i'm i'm so grateful to you for making more people aware of this fact that humanity has reached the most important fork in the road ever in its history and let's turn in the correct direction thanks for listening to this conversation with max tagmark to support this podcast please check out our sponsors in the description and now let me leave you with some words from frank harbert history is a constant race between invention and catastrophe thank you for listening and hope to see you next time