TRANSCRIPT: Connecting AI to the internet is a big mistake | Max Tegmark and Lex Fridman

Intro

if you're somebody like Sandra pachai 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 if it's if it's created very very soon and as 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 you know Sam Altman and then Mr sabis have both said themselves acknowledge that you know there's really great risks with this and they they want to slow down once they feel it gets scary it's but it's clear that they're stuck in this again molok 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 uh it's just to win this wisdom race

The wisdom race

it's clear that what we hope that is going to happen hasn't happened the the capability progress has gone faster than a lot of people thought than and the part the progress in in the public sphere of policy making 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 um large language models and other poorly understood systems couldn't get us all the way that 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 um I'm quite confident that this can be done um so we can reap all the benefits but we cannot do it as quickly as uh this is out of control Express train we're on now is gonna get the AGI that's why we need a little more time I feel is there something to be said well 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 uh don't invest in a closed development where you focus on AI safety while is somewhat dumb quote unquote uh release as often as possible and as you start to see signs of uh human level intelligence or superhuman level intelligence then you put a halt on it well 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 uh 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 yukowski you said you interviewed him recently right yes yes so he had this tweet recently which I God gave me one of the best laughs in a while where he's like hey people used to make fun of me and say you're so stupid Eliezer because you're saying you're saying um you have to worry of 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 yeah where you know you can really study it if so he had written it in the like in the meme form so it's like then yeah and then that and then now LOL let's make a chatbot yeah yeah and the third thing is Stuart Russell yeah you know amazing AI researcher he

AI and humans

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 how 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 recommender algorithms which do exactly that they they get so good at knowing us and pressing our buttons that we've we're starting to create a world now where we just have 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 the euphemism get people glued to their little rectangles right it's 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 non-stop signal it'll show a thing and it will give signal on whether we spread that thing we like that thing that thing increases our engagement gets us to return to the platform and 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 param the number of parameters in your neural network that's doing the learning and more end 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 you think is in Humanity's interest yeah 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 ai's persuade humans to let them Escape whatever safety precautions we put you know there was a really nice article um and the New York Times recently by you've all know a Harari and 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 on 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 then people can't even agree on who won the last election you know

Social media

and we humans often point fingers at other humans and say it's their fault but it's really moloch and these AI algorithms we got the algorithms and then molok pitted the social media companies are against each other so nobody could have a less creepy algorithm because then they would lose out on our 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 is it possible uh to do social media correct I think it's not only possible but it's 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 it'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 that's it's yeah we both agree let's not have a nuclear Wars let's not do that um etc etc we're kidding ourselves thinking we can face off the second contact with whatever more powerful AI that's happening now with this large language models if we can't even have a functional conversation in the public space that's why I started to improve the news project improve the news.org but um

What makes the difference

I I'm an optimist fundamentally in um and that there is a lot of intrinsic goodness in in people and that uh what makes the difference between someone doing good things for for Humanity and bad things is not some sort of fairy tale thing that this person was born with the evil Gene and this one is 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 they 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 they both make money and bring out the best in people I mean in the long term it's not a good investment for anyone you know to have a nuclear war for example and you know is it a good investment for Humanity if we just ultimately replace all humans by machines and then we're so obsolete that eventually the there are no humans left well it depends against how you do the math but like if I would say by any reasonable it cannot be started if you look at the future income of humans and there aren't any you know that's not a good investment moreover like why why can't we have a little bit of pride in our species damn it 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 the if we had really Advanced biotech to build homo sapiens you you know you might say hey Max you know yeah let's build build the these Homo sapiens they're going to be smarter than us maybe they can help us defend us better against the Predators and help fix up our caves make them nicer and we'll control them undoubtedly you know so then they build build a couple a little baby girl little baby boy you know and and then you have some some wise old and the enderthal Elder was like hmm I'm scared that uh we're opening in 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 are you such a Luddite scaremonger or of course they're going to want to keep us around because we are their creators and and why you know the smart I think the smarter they get the nicer they're gonna get they're gonna leave us they're gonna they're gonna want this around and it's going to 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 gpt4 yeah it's not I want to be clear it's not gpt4 that's terrifying it's the gpt4 is a baby technology you know at Microsoft even had a paper recently out the 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 they'll be way more powerful and but they're going to take all the things ideas from these babies and before we know it we're gonna be like those last neanderthals who were pretty disappointed and when they realized that they were getting replaced well this interesting point you make which is the programming is it's entirely possible that GPT 4 is already the kind of system that can change everything by writing programs sorry it's yeah it's because it's Life 2.0 the systems I'm afraid of are going to look nothing like a large language model and they're not 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 you know the dumbest way that still works sort of yeah and um so they are live 3.0 except when they replace their software it's a lot faster than when you when when you decide to learn Swedish and moreover they think a lot faster than us too so when uh you know we don't think uh have one logical step every nanosecond or a few or so the way they do and we can't also just suddenly scale up our Hardware massively in the cloud because we're so limited right so they are and they are also life have consumed become a little bit more like life 3.0 and 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