

[Max Tegmark explains how quickly time is running out to manage the risks posed by AI \(7/8\)](#)

OxfordUnion. 17,868 views. 27 Nov 2023.

It's such an honor and pleasure to be here in England's greatest University. I've counted many different definitions of of existential threat. Tonight if we were in some very sloppy thinking place like Cambridge I would just leave it at that. But because we're in Oxford I'm going to define what I mean by extensional threat. First of all the term existential risks was defined by an Oxford professor and I'm going to go with his definition which is a threat of human extinction or permanent permanent limiting of ohuman potential. And that'll become important in a little bit. What about threat? What is a threat exactly? Well, for example Silva you defined a threat as something which is 100% certain to happen and then very successfully attack that. But I looked it up in the, not in the Cambridge Dictionary, in the Oxford dictionary, and it said that a threat is a possibility of something bad happening. What does a possibility mean? It means the probability of the bad thing happening doesn't mean it's 100% it means it's more than zero. And we had a good question from up there somewhere. Where do you draw the line? Well if you're talking about something like extinction suppose there is an asteroid which will kill us all if it hits us? What is the probability threshold where you feel you really need to take action? You have to look into your hearts and answer that. Is it one in 10,000? Is it 1 in a 1000? One in 100? Regardless of what it is that's what you should think of when you choose which door to walk out of. If you pick 1% if you think the risk of an existential threat is more than 1% you go through. Then you're yay now so that's that now we got the definitions out of the way. Except there were two really interest there was another interesting twist also two of the opposition speakers said it's only a threat if it's happening right now and they said said it's AI is not a threat because we don't have AGI yet super intelligence. So suppose this asteroid that's heading towards Earth again isn't going to hit us for 20 years. Does it make any sense to say it's not a threat even if the best astronomers have calculated it has a 99% chance of hitting us? That's ridiculous. And ironically Sebastian when you said that the you only count risks that are happening things things are happening now not in the future is they turned around and said that the

catastrophic climate change is a threat even though it's happening in the future. So now (yeah ask me in in in four minutes) so okay so why should we worry about this stuff? Then raise your hand if you study history. You know then, that we can only really understand the present if we study history. I think it's very easy to get so blinded by all the recent stuff that's happening with AI that we lose the context. So if you are lucky enough to have 50 Quid in your pocket can you take it out please for a moment. Let's see how how well -Oxford is charging too high fees I think - so this gentleman here Alan Turing has already been mentioned really the Einstein of computer science, an amazing man, not only viewed as the founding father of computer science and of artificial intelligence also really an amazing war hero. Raise your hand if you saw imitation game the movie. Yeah well then I'm preaching to the choir here [he] helped crack the Germans code by building the first ever programmable digital computer. Pretty smart dude. Sso he said back in 1951 as Jaan already mentioned that if we ever get to machines that can outsmart us then they will pretty quickly leave our feeble intellects far behind and we should expect he said that at some point they will take control. The machines will take control. That's not particularly rocket science. When we showed up in the rainforest tough luck for a lot of the animals that lived there. We took control, not because we had bigger muscles but because we were smarter and it didn't always work out for them. But he had, this is his argument, and it was we have to pay respect to his intellect, we can't just diss him without giving good counter arguments. I know it sounds like science fiction machines, taking control, but you know nuclear weapons sounded like science fiction in 1930 and when Alan Turing formulated the Turing Test which he heralded as the final Milestone before we should really start freaking out. Having something like ChatGPT-4 sounded like science fiction also. So saying that something sounds like sci-fi does not imply it won't happen- that's just a really really lame argument. And it's not just Turing. Sam Alman CEO of OpenAI that's given us ChatGPT has said himself recently that it could be "lights out for all of us". That's an existential threat isn't it? And in May of this of this yar, and in fact Dario Amado, CEO of Anthropic, one of the AGI competitors even put a number on this risk. He said he thinks we have two or three years left

until we get outsmarted. (information, in 3 minutes I'll take it) And it's not just them it's not just them. In May of this year we had all the CEOs of the AGI companies, Google Deep Mind here in London, OpenAI, Anthropic and who's who of leading AI researchers, including at Oxford, say this is an existential threat. Human extinction is something we have to take seriously. (So I said two minutes you you have, you seem to have four clocks here are you from Cambridge?) So, this sounds kind of dark. So let's have, let's raise our spirits a little bit by having some fun and playing Jann Tallinn bingo. So he told us that we should keep a bingo card of the opposition argument so that's actually what I've been doing here. He said there are four kinds of bad arguments that tend to be made. Arguing that it's not a threat. One of them is labeling atomm attacks person X said y also which therefore X is wrong and and um K sh I agree with you on many things but you got that little Bingo score right there because you said Nick Bostrom warned of existential it's an existential threat and he said these really bad things therefore it's not an existential threat so yeah so got you and the framing so we have many winners here so Sebastian you said oh AI is not evil therefore it's not an existential threat technology isn't evil and it's also not morally good it's a tool and we can use it for good things and bad things. Turing was well aware of this and that does not change the fact that it's an existential threat. Sebastian also said that it's yeah actually let me pick on Silva for some democracy here actually let's move on I picked on you once already I want to spread my graces. Human Supremacy. We had the argument that AI cannot compete with us no matter how advanced it gets because it doesn't have quite our human culture and the art isn't as profound. This was discussed here. Right, well you know sorry to say Africa had some amazing art when British and other colonialists came there with big guns and that did not protect these people in Africa at all. Something doesn't have to be more cultured than us to pose a threat. In fact if you think about Alan Turing's warning why he was warning if you think of AI is just another technology like the steam engine or electricity you won't be very worried. But you see Alan Turing was taking the Long View he was thinking of it more as a new species. We are, as a matter of fact, right now building creepy super capable amoral Psychopaths that never sleep, think much faster than us, can

make copies of themselves and have nothing human about them whatsoever. What could possibly go wrong? Once we have out there in the world entities that we cannot control I'll talk about how to control them tomorrow if you come back at 5:00 P p.m. But if if we can't control them and uh they are much better at us at thinking, speaking, manipulating people, getting things done, the future belongs to them not to us. This is what Turing realized and then the 72 years since nobody has convincingly refuted that argument. It's really as simple as that. We stand to lose control. And to close, an existential threat, according to the Oxford definition, is not just something that drives us extinct it can also mean something that permanently limits our potential. So if we lose control to machines either be directly to the machines or to some horrible people through too much power concentration that managed to control us forever in some sort of dystopia 1984+ You know that is an existential threat by this very definition. So let's not do this and I want to just end with a note of Hope. Jaan you gave them all a term diagnosis. That's a little too harsh. I think there is still might this might still be treatable. So what I think you're really voting on if you vote Yes for existential threat is vote Yes for taking this really seriously and being really motivated to go out there and do something about it. Thank you.