**Will Super-intelligent AI End the World?**

Eliezer Yudkowsky

Since 2001, I have been working on what we would now call the problem of aligning artificial general intelligence: how to shape the preferences and behavior of a powerful artificial mind such that it does not kill everyone.

00:19

I, more or less, founded the field two decades ago, when nobody else considered it rewarding enough to work on. I tried to get this very important project started early so we'd be in less of a drastic rush later. I consider myself to have failed.

00:34

Nobody understands how modern AI systems do what they do. They are giant, inscrutable matrices of floating point numbers that we nudge in the direction of better performance until they inexplicably start working. At some point, the companies rushing headlong to scale AI will cough out something that's smarter than humanity. Nobody knows how to calculate when that will happen. My wild guess is that it will happen after zero to two more breakthroughs the size of transformers.

00:59

What happens if we build something smarter than us that we understand that poorly? Some people find it obvious that building something smarter than us

that we don't understand might go badly. Others come in with a very wide range of hopeful thoughts about how it might possibly go well.

01:16

Even if I had 20 minutes for this talk and months to prepare it, I would not be able to refute all the ways people find to imagine that things might go well.

But I will say that there is no standard scientific consensus for how things will go well. There is no hope that has been widely persuasive and stood up to skeptical examination. There is nothing resembling a real engineering plan for us surviving that I could critique. This is not a good place in which to find ourselves.

01:44

If I had more time, I'd try to tell you about the predictable reasons why the current paradigm will not work to build a superintelligence that likes you or is friends with you, or that just follows orders. Why, if you press "thumbs up" when humans think that things went right or "thumbs down" when another AI system thinks that they went wrong, you do not get a mind that wants nice things in a way that generalizes well outside the training distribution

to where the AI is smarter than the trainers. You can search for "Yudkowsky list of lethalities" for more.

02:22

But to worry, you do not need to believe me about exact predictions of exact disasters. You just need to expect that things are not going to work great

on the first really serious, really critical try because an AI system smart enough to be truly dangerous was meaningfully different from AI systems stupider than that. My prediction is that this ends up with us facing down something smarter than us that does not want what we want, that does not want anything we recognize as valuable or meaningful.

02:52

I cannot predict exactly how a conflict between humanity and a smarter AI would go for the same reason I can't predict exactly how you would lose a chess game to one of the current top AI chess programs, let's say Stockfish.

If I could predict exactly where Stockfish could move, I could play chess that well myself. I can't predict exactly how you'll lose to Stockfish, but I can predict who wins the game. I do not expect something actually smart to attack us with marching robot armies with glowing red eyes where there could be a fun movie about us fighting them. I expect an actually smarter and uncaring entity will figure out strategies and technologies that can kill us quickly and reliably and then kill us.

03:34

I am not saying that the problem of aligning superintelligence is unsolvable in principle. I expect we could figure it out with unlimited time and unlimited retries, which the usual process of science assumes that we have. The problem here is the part where we don't get to say, “Ha ha, whoops, that sure didn’t work. That clever idea that used to work on earlier systems sure broke down when the AI got smarter, smarter than us.” We do not get to learn from our mistakes and try again because everyone is already dead.

04:07

It is a large ask to get an unprecedented scientific and engineering challenge

correct on the first critical try. Humanity is not approaching this issue with remotely the level of seriousness that would be required. Some of the people leading these efforts have spent the last decade not denying that creating a superintelligence might kill everyone, but joking about it.

04:30

We are very far behind. This is not a gap we can overcome in six months, given a six-month moratorium. If we actually try to do this in real life,

we are all going to die. People say to me at this point, what's your ask? I do not have any realistic plan, which is why I spent the last two decades trying and failing to end up anywhere but here. My best bad take is that we need an international coalition banning large AI training runs, including extreme and extraordinary measures to have that ban be actually and universally effective,

like tracking all GPU sales, monitoring all the data centers, being willing to risk a shooting conflict between nations in order to destroy an unmonitored data center in a non-signatory country.