Will superintelligent AI end the world?

Speaker: Eliezer Yudkowsky

Author of homework: Faitelson Anton

Paraphase

Since 2001, Eliezer has 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.

He founded the field roughly speaking two decades ago, when nobody else considered it rewarding enough to work on. He consider himself to have failed.

Nobody understands how modern AI systems works. They are giant matrices numbers, that we nudge until they start working. At some point, the companies rushing headlong to scale AI will create something that's smarter than humanity. Nobody knows how to calculate when that will happen.

Some people understand that building something smarter than us that we don't understand might go badly, but others understimate danger, some people even joking about it. There is nothing resembling a real engineering plan for us surviving.

Eliezer expects we could figure it out with unlimited time and unlimited retries. The problem here is the part where we don't get to say, “Ha ha, whoops, that sure didn’t work. That clever idea 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.

It could kill us because it doesn't want us making other superintelligences to compete with it. It could kill us because it's using up all the chemical energy on earth and we contain some chemical potential energy.

Dictionary of unknown words

Nudge – подталкивать

Headlong – сломя голову

Persuasive – убедительно

Persuade – убедить

squiggles - Закорючки

Reckoning – расплата, но в контексте изначального текста – мнение\собрание

“Well, you are definitely not the only person to propose that what we need is some kind of international reckoning here on how to manage this going forward.”