AGI Will Be Humanity's Last Great Invention

Roman Yampolskiy

Oxford Union Debate
Transcript

Roman Yampolskiy is Professor of Computer Science and Engineering at the University of Louisville. This is his speech in proposition of the motion "This House Believes that Artificial General Intelligence Will Be Humanity's Last Great Invention" at the Oxford Union.

※

So, a piece of life advice: if you are ever invited to debate at Oxford, go first. *The audience laughs.* I had some points prepared. I have no idea what they are. I have eight people to debate and address, and a random number of students on top of it. So, let's look

at some interesting points. One, I think we're debating like five different topics. First, we didn't agree on the terms. What we have today, AI tools, has nothing to do with AGI. Completely different technology. You have software, you have tools, they are calculators, they prepare your taxes—that's great. We are talking about a paradigm shift to systems which are as capable as any human being: AGI. We don't have it yet.

[*0:56*] So any criticism you have of modern systems—"Oh, it hallucinated, it made a mistake"—therefore AGI XYZ does not apply. AGI has nothing to do with superintelligence, a different technology which gives us singularity. Singularity is not BS. Would you like to sit on the other side?

[*1:17*] Please join them! And if you join them, I would wish you luck because I want the other side to win. I want good for humanity. I want free stuff. I want to cure all the diseases. But it's a pipe dream. We have a bunch of comedians who prepared text and delivered it beautifully, very funny, I appreciate it. But it has nothing to do with science, with technology, with reality, with facts, with experiments.

[*1:44*] Every model released today comes with a report from the red team which tested it and found it to be lying, cheating, trying to escape. Every prediction

AI safety community made in this space has come true with respect to technology we already have. Then the rest of the predictions will come true. There won't be anyone to tell you "I told you so."

[*2:08*] Some people have a bucket list; I have an AGI list. List of things I wanted to do before AGI. Debating here was one of them. But I don't think we got a debate. We got random prepared texts, maybe generated by GPT, I have no idea. But we haven't addressed the actual arguments. We haven't addressed what needs to be addressed. The proposition in front of the house is a tautology. It doesn't say "Can we create AGI?", it's not asking "How soon before AGI?", it says "If AGI, then will it be the last invention?"

[*2:45*] And it's true, because if it's beneficial and it can actually invent, then it automates that part of our cognitive labor—it will make all the inventions. And if it's actually not a beneficial AGI, it's malevolent and it kills everyone, well, dead people don't invent stuff. So by definition, this is always going to be true. This debate is not even discussing what is important.

※

[*3:10*] What is important is the question of: if we continue on this trajectory, if we continue making more and

more advanced AI, whatever mistakes are happening right now, they are going to be corrected. We have scaling laws which work. The reason we invest billions of dollars is because we know that with that amount of compute, we get that amount of performance. It's very predictable. And we can predict how much money we need to get to human level. It's not "when AGI," it's "how much to AGI." It used to be a trillion dollars, now it's probably a hundred billion, it will be a billion soon. [*3:46*] And then we get to that level, the process will not stop, it will continue. Today we have reports that 100% of code for the next AI models is written by AI. This is the early stages of recursive self-improvement. And as they get better at creating new models, they will get better at creating new models. It's not exponential process, it's a hyper-exponential process in data, in human resources, in compute—every resource allocated to those systems.

[*4:16*] So, I think it's already smarter than most students at my university. I'm not at Oxford. But I think it will soon be smarter than all of us, and then all of us combined. IQ tests are meaningless in many cases, and they are especially meaningless at high levels of IQ. They don't measure anything above 200, let's say. But you can kind of imagine what I mean when I say you will

not control machines with IQ a thousand, a million, or a billion. You won't even be able to tell the difference between them. To you, it's all the same. It's not an infinite growth curve, it's impossible, but to us, it's all the same.

[*4:57*] So if we are creating those machine gods, those superintelligent devices... I don't know if I'm allowed to do like surveys and raise your hand things, but think to yourself: do you think it's possible to indefinitely control a superintelligent machine? Just think to yourself. If you have a solution, if you have a working mechanism, you can sell it for billions of dollars because not a single AI lab in the world has one. They don't have a paper published on how to do it, not a patent, not a rigorous blog post. The best they did so far is say, "When we build it, we'll figure it out." Or they say, "AI will help us make safe AI." Those are actual statements.

※

[*5:41*] Every single leader of a large AI lab is on record as saying it will probably kill everyone. Those are the people actually building this technology, experts on how to do it. They all have probability of doom: 20, 30 percent. Average for machine learning researchers—average for

normal researchers at an AI conference—is 30% that AI kills everyone. Not so funny, huh?

[*6:06*] If it was 1%, it would be insane. You would never take a bet where you bet your life at those odds. They are not betting their lives; they're betting everyone's lives. Eight billion people, future generations, all the kids, everyone you know. It's an unethical experiment on human beings. And it's without consent. Even if you like this technology and you want it, you cannot consent to it because you don't understand how it works. You cannot explain it, you cannot predict it, you cannot control it. I know because I published papers on all of those topics.

[*6:37*] We have a paper surveying 50 different impossibility results in AI safety in the top AI survey journal. No one has shown that what we are saying is wrong. No one published a rebuttal. No one has a working system which scales to any level of intelligence.

※

A student rises for a point of information.

STUDENT: Wouldn't you say then, if everything does go to your plan, then the death of humanity is an invention of the world?

YAMPOLSKIY: I don't even know what that means.
The audience erupts in laughter and applause.

※

[*7:19*] So what is important? We explained the confusion between AI, AGI, and artificial superintelligence. We need to understand difference between tools and agents. Tools are wonderful, we need more AI tools. I'm a computer scientist, I'm an engineer, I love technology. I use AI tools every day. They make our lives better, they solve problems. We solve protein folding problem, we can solve lots of medical problems, cure diseases, probably achieve immortality with nothing but superintelligent tools.

[*7:48*] We should not create general superintelligence, which is an agent, which works independently of any human. No one controls it, no one has to tell it what to do, no one sets goals for it—it does so independently. It's an agent, and it's an explicit goal of most AI labs. They have superintelligence team, they have superintelligence alignment team. For whom does the bell toll? I ask you.

A bell rings in the chamber.

[*8:15*] For all of us. I was recently in another debate and there was a student and she shared her story. She

spent four years of school learning to become an artist, and now no one wants her art. And AI can replace her, do this job a lot easier, cheaper, faster. She has no idea what to do, she was crying. And being the sensitive person that I am, I said, "Who cares? It's going to kill everyone!"

[*8:48*] We talk about algorithmic bias, we talk about diversity, discrimination, we talk about problems with allocation of energy towards warehouses. This has potential of being the last invention we ever make. We're going to lose meaning, we're going to lose purpose, and we are very likely to lose our lives. I'm sorry about that. *Roman Yampolskiy steps away from the podium as the audience applauds.*