AI and You

Transcript

Guest: Roman Yampolskiy

Episode 16

First Aired: Monday, October 5, 2020

Hi, welcome to episode 16. Today we have a feast, because I am talking with Dr. Roman Yampolskiy, a legend within the AI community. He is a tenured associate professor in the department of Computer Engineering and Computer Science at the University of Louisville in Kentucky where he is the director of Cyber Security. Look up his home page and Wikipedia entry because he has qualifications too numerous to detail here. Some of those include that he is also a Visiting Fellow at the Singularity Institute, he was also one of the scientists at the Asilomar conference on the future of AI, and the author of over 100 publications, including the book *Artificial Superintelligence: A Futuristic Approach*, which I would strongly recommend, and many peer-reviewed papers, including one coauthored with yours truly. Most of them are on the subject of AI Safety.

In this episode we will talk mostly about a recent paper of his, a table-thumping 73 pages titled "On Controllability of Artificial Intelligence," which you may recognize is talking about what we call the *Control Problem*. That's the most fundamental problem in the whole study of the long-term impact of AI, and Roman explains it very well in this episode.

If you're looking at the possible future of artificial intelligence having serious negative as well as positive consequences for everyone, which is a concept that's gotten a lot of popularization in the last few years, of course, and you're finding computer professionals pooh-poohing that idea because AI right now requires a great deal of specialized effort on their part to do even narrow tasks, and the idea of it becoming an existential threat is just too sensationalist, the product of watching too many Terminator movies, then pay attention to Roman, because he has impeccable computer science credentials and yet does not shy away from exploring fundamental threats of AI. In fact I would say the number of people fitting that description is increasing. Stuart Russell at Berkeley comes to mind as another example.

With so much to talk about it's no wonder that this interview is spanning two episodes again. In this part of the interview I reference the *Chinese Room* argument and say I'll explain it in the show opening. Actually, I gave a lengthy explanation of the Chinese Room at the beginning of the last episode and here we are again, but rather than force you to go back to episode 15 if you *haven't* heard it, I'll repeat it here. If you've just listened to episode 15 – or if you already know what the Chinese Room is – you can skip ahead a couple of minutes.

The Chinese Room is the name of a thought experiment by the philosopher John Searle, from the University of California at Berkeley. You may know that the Turing Test, what Alan Turing called the *Imitation Game,* is an empirical test to determine whether or not we should consider that a computer is thinking like a human. If, in a blind test of its behavior we cannot tell that it's not human, then Turing said we should decide that the computer is thinking like one. We should talk about the Turing Test in more detail some time, but that's enough for now.

Searle does not like Turing's conclusion. So he came up with a rebuttal that goes like this: Imagine a room. We're only seeing it from the outside, where it looks like a shipping container, completely sealed, except for a hole on one side and a hole on the other. Into one hole we put a piece of paper with a question written in Chinese on it. Some time afterwards, a piece of paper comes out of the other hole with the answer, also written in Chinese on it. Searle's question is, does the room understand Chinese? And then he goes on to say that inside the room is a man and a book. The man does not understand Chinese at all. But the book contains instructions that tell him what to do with the pieces of paper that come through the input hole. "If you see a mark with this pattern, then go to page 572 of the book and read rule 12 there. If it's followed by a mark like this, go to such-and-such page. Write a mark that looks like this on the output paper." And on and on, for many, many instructions. We know that this is a reasonable scenario because computers can answer questions put to them in Chinese by following instructions that are like that. You just need a big enough book and a lot of time, but since it's a philosophical thought experiment, the size of the book and the amount of time are not barriers. Searle says that since the man does not understand Chinese, the room is not understanding Chinese, and that therefore if a computer gives the appearance of answering a question put to it in Chinese it doesn't understand Chinese either, and the same goes even if the question is in English. Therefore, computers can't understand anything and they can't think like humans.

If you think that Searle pulled a fast one along the way there, you're not alone. There are many rebuttals to his argument, the most common one being that the *system* of the room *plus* the man *plus* the book does in fact understand Chinese. Furthermore, by using objects that to us have obvious limitations in the real world – books have hundreds of pages, not billions, and people have neither the patience nor the time to follow a trillion rules in such a book to solve a problem like this – he is being so reductionist that we lose sight of the possibility of *emergent behavior* such as *understanding* in a system that complex. Yet a system that complex is *exactly* what we know the human brain is. Similar arguments are made for claiming that evolution could not produce organisms as complex as the ones on Earth right now, but those are made by people who cannot appreciate just how long a billion years is. To be fair, none of us can really appreciate how long a billion years is, but scientists are at least aware of their own limitations in that regard.

A great deal of academic ink has been spilled pointing out where Searle went wrong, mostly because he has never given up defending his argument. In that respect we should at least be grateful that he's provoked so many people to critical thinking by giving them such a large target to throw rhetorical darts at. You might want to try your hand at that too.

Another thing we reference in this interview is GPT-3. That stands for Generative Pre-trained Transformer – you can see why everyone just says GPT – and this is the third version of it. GPT-2 ingested a large amount of text from the Internet and was able to interpret and construct statements in natural language to a surprising degree; but it had 1.6 billion parameters whereas GPT-3 has 175 billion parameters and it is really impressive. Close to Turing Test passing grade. If you google for it, you'll find places on the web where you can see what it has done and even interact with it yourself.

Okay, enough from me, let's get on with part one of the interview with Roman Yampolskiy.

Roman, welcome to the show.

Thank you.

How is everything going at the university there with classes starting and the coronavirus? How is the morale among the faculty?

It's different. We went from teaching in person, face to face to teaching mask to mask or online or hybrid. Essentially, we are good for anything. If [an] asteroid hits us, we're still prepared.

Oh, wow. I haven't seen any papers from you yet about asteroids hitting us but I guess at your rate of production, that might be in the works. Can you tell us how did you get into this field? What sort of events and thinking in your life prompted this line of work?

So I was doing a lot of work on games, online poker and trying to profile players based on their behavior. A subfield of behavioral biometrics where a game strategy is used to identify let's say, stolen accounts, so behavior changes, we can assume someone else took over your account. And I realized so many players online were not human, they were bots. So I started trying to detect them, prevent them, and then I realized the bots will get smarter. And this is where we are today - AI safety and security for superintelligent bots.

Right. And let's focus on that superintelligence. You've got a recent paper "On Controllability of Artificial Intelligence"; it's 73 pages. That's large even by those standards. I'd like to talk about that in this podcast. A lot of it is over my head, but we'll talk around it and how it relates to people that are listening to this podcast especially, and people in general because while a lot of academics in this field keep to themselves, stay in the ivory tower, you make a point of getting out there and talking to laypeople. What's your philosophy on that?

Well, we're going to impact their lives. I think they deserve to know what we're doing for them and to them and maybe they have some opinions they want to share with us and help us decide what to do and not to do. Seems like if your science is not public, are you really doing it?

Thank you. Now, this paper addresses what's called in the field the control problem and I think in the conclusion, you've got the most succinct summary I've seen of the control problem in AI yet, and the implications of it. Can you go over that for us, please?

All right. So the problem is as AI becomes more capable, we go from narrow AIs tools systems to play chess, be calculators, to AI which is as smart as humans and eventually smarter than humans, can we remain relevant? Can we remain in control? It doesn't matter what architecture, what algorithms are used, you can have multiple definitions of control. You can talk about direct control - I give orders, you follow them. We can talk about some sort of implied control where you're trying to figure out what I want and make my life easier. And in the paper I try to look at all those different definitions and see well, "Under any definition can we get what we want from a superintelligent machine?" And it looks like for each one of those types, there are side effects. Problems with direct control, you get kind of a standard genie problem - you have three wishes and you always regret making any of them. With implied control, the problem is, what do you contribute to superintelligence? Why does it need you? If it's smarter than you, if it's deciding what needs to happen, why are you even part of this equation?

And that's touching on philosophical angles here, and I see that the paper is published in philosophical paper archive, as opposed to computer science. Now, you come from a computer security background. In fact, you're, I think, director of a computer security lab, is that right?

That's right. So those are different archiving services. In fact, you correctly complained that 73 pages is a little too long, so like good McDonald's, I now make this paper available in small, medium and large in different archiving services, including computer science. I have a shorter version. I think I got a 36-page one and 59-page one. So depending on who you are and what discipline you're in, again, I'm trying to reach out to every single possible reader.

I see. So I'm looking at the supersized version.

I trust you to handle the Big Meal.

Yeah, thank you. Not exactly a Happy Meal. Now, you are in computer security. I work with computer security people every day. They're focused on threat detection. They're focused on intrusion alerting, traffic analysis, information security classification. AI is not on their radar. Should it be?

So that's the security component. I talk about AI safety and security. They are looking at the security portion, the external threat to the system - hackers getting access to the code, changing it. But if a system itself becomes an agent, now you have [an] insider threat. The system is possibly dangerous to its users. So that's the safety component.

You talk about it becoming an agent. And that gets at something fundamental I've been poking at with a number of people that I've been talking to, which is that right now, people don't talk about the AI we have at the moment having agency. They don't ascribe motivation or values to it unless they are stretching a point. But in discussions like the one that you have about artificial superintelligence, we do talk about it having some kind of motivation. We do talk about it having agency. We talk about it having goals. At what point would AI cross the line where we are ascribing that kind of behavior to it?

So right now we don't have general intelligences. We don't have human-level or superintelligence. So we deal with narrow systems and they are not agents, they are tools. Once we get to that level of human equivalence, I think it's fair to call them agents.

Could we say that any program that doesn't do what we intend it to do is a reductive example of motivation in some sense?

Well, again, with narrow systems, if they don't do what we want, it's a mistake in coding. It's a bug, it's not necessarily some sort of intent we can attribute to the system. We've failed to properly tell it what to do.

But when we have an agent, an AI that we ascribe agency to, and it's taking some independent action, we ascribe that agency by virtue of the fact that it's doing things we didn't predict, that

weren't on the list of things we told it to do. Is that an example of a bug writ large, a program not doing what we intended it to?

Well, with agents and human-level performance, we cannot specify all possible outcomes. It's like having children. You can hope for good children but you really don't know what they're going to become. And it's pretty much the same. They are independent agents - you are not responsible for what they do because you cannot be.

Well, that's a good example there to talk about children. I've got a 10-year-old and a six-year-old and it's being rubbed in my face daily that any idea that I have of controlling them is rapidly becoming an illusion if it weren't always the case. And yet, we don't worry about our children growing up to destroy the world. Although if you saw the way mine played together, you might realize that's not a foregone conclusion but still, we don't worry about that. Now, if we consider AI to be children that will one day grow up in some sense, why should we worry about what they will grow into?

Well, I've got three kids and I very much worry about them destroying the world. I think the difference is power. Your children are, I assume, average human beings, so they have [the] power of an average human being. A superintelligent system is more like a God agent. It has superpowers compared to an average human so the level of destruction will be proportionate. If it's not well aligned with our goals, it can mess things up. If, not your kids, but if some human becomes a crazy psychopath serial killer, they'll get 10 people, kill 20 people, maybe start a war, but they're unlikely to wipe out a billion humans, whereas a much more powerful system can.

Right. Well, and this is perhaps a different topic, although one that I explored in my book, but as we develop more capable technology and say, biogenetics, then people will have the capability to wipe out a billion people and so perhaps people will also become the superintelligences of the future that we should worry about.

Absolutely. We are not limited by substrate. We're concerned about weaponized intelligence, greater intelligence. If we can do biological superintelligence, it's the same set of problems. Mind uploading is another possible path to that.

Could we consider aggregations of human behavior right now like governments, armies, and corporations to be superintelligences?

In a certain way - they're greater than any given human by itself but I think they're not that far from a smaller group of people, for example, so 100 people versus 10,000 people. I think the difference would not be that tremendous, whereas again, with super intelligence, the upper bound is unknown to us.

Let's look at some of the commentary around this topic. It's certainly controversial in that, as seems to happen anytime in science where there's not enough information to prove something one way or the other, the scientists separate into two opposing camps that hate each other's guts. And the controllability of AI has its detractors, and you have people like Andrew Ng saying that worrying about AI is like worrying about overpopulation on Mars. Now, technically, he's

not saying it's impossible because far enough in the future, Mars may well become overpopulated and we would have to worry about it, but he's saying it's not worth worrying about that now; there are other things to do. Now, if we take the word worrying out of the equation, because while people do get alarmed about this sort of thing, your paper is academic, it's clearly something that you have decided is worth you spending time thinking about right now. How much should the rest of us be thinking about this? How much attention should we be devoting to it? And perhaps the precautionary principle should be invoked here as some kind of justification.

Well, we don't know how soon it will happen, right? I heard estimates as optimistic or pessimistic, it depends on your view, as seven years, five years, and some people say hundreds of years, thousands of years, but [the] problem is the same. The difficulty is the same. Impact is the same. So really, the question is, should we worry about problems far in advance if we think they're not coming soon? You can look at things like [the] invention of vehicles. Initially, cars were electric. Because people didn't think about climate change and oil consumption, they switched to gas vehicles. If they spent some time a hundred years ago on this problem, we would have less problems today. So even if predictions about [the] overpopulation of Mars [is] used as an example, we can do something today to make problems for future generations easier. Why not do that? Lots of people work on immediate AI safety concerns - face recognition, unemployment, bias in AI - I'm happy they do. It's wonderful. But if we don't allocate a certain number of people to bigger problems, which may be more long-term problems, then we're not optimizing our future in any way I'd like to see it.

Got it. Now, a lot of people that work in the field of AI right now have some disdain for this point of view. There's an adaptation of a popular meme, the one with the screaming women and the cat on the other side at the dinner table. And on the left side, the caption over the screaming women is "People who are thinking AI is going to take over the world" and on the right-hand side, the caption is "My neural network classifier", and there's a bounding box around the cat's head with a tag that says "Dog". And that's accurate in the sense that AI right now, people have to do an enormous amount of work to do even simple image classification and that's fragile with respect to the kind of alterations that don't impede humans in the slightest. Does that level of discrepancy, that level of difference between where we are right now and what you're talking about get you any funny looks from people that you're, say teaching C++ to or down in the trenches with computer science?

I love memes. If you think about that specific one, they use example of AI failing at its only job as a reason not to worry about AI safety. They couldn't even come up with an AI which doesn't fail today and it's only going to get worse. People who give you funny looks are the ones who never even for a second considered what happens if they succeed. They don't think they will succeed. They see it as some sort of perpetual job. We'll always try to increase accuracy, but we'll never get there. So in general, I'm not worried about opinions of other people and much less about those who don't actually think about deep future.

The paper talks about *Friendly AI,* which is a term that's been around for a while. Can you explain that?

So there has been a lot of terminology trying to address the same concept, [whether] it is well-aligned artificial intelligence aligned with human values, Friendly AI, Safe AI, which is kind of going through different terms, but the meaning is, we want beneficial intelligence systems which give us something we will be happy with, not harm us.

Right. Now, you quote Eliezer Yudkowsky as saying that, "The assertion that it's impossible to make friendly AI - AI that won't hurt us - is in the same class of statements like it's impossible to fly to the moon." That seems to me to be inverting the reality. Surely the statement should be "It's impossible to prevent unfriendly AI because no matter how many friendly AIs we make, it only takes one unfriendly AI to upset the applecart." Would you agree with that assessment?

All right, so in the paper, I try to give voice to every side of the issue - opinions showing it's impossible, it's possible, different degrees of impossibility. I do argue in a number of papers about malevolent development, malevolent actors that you cannot prevent that. You can fix bugs, you can fix mistakes, but if someone on purpose on the inside of a team tries to cause some damage, it's very hard to do anything about. I don't think I've found any solutions for that particular issue yet.

Well, speaking about malevolent actions, you were part of the Asilomar conference that came up with the 23 Principles for Ethical Development of AI, including some long-term ones that directly address the kind of issues that you're raising in your paper here. Should those principles see some more adoption? Should we see, like, Google having a stamp of "Asilomar Compliant" or something like that?

I don't think there is shortage of principles or ethical codes. I think everyone who's anyone has released a set of such principles, [whether] it's United Nations, World Forum, everyone. I think latest survey paper looked at like 50 sets of ethical guidelines and standards, and it might as well be 5000. It doesn't matter unless you're actually doing something practical with those things. It's like mission statement for a university. Someone reads those, we just released a new one. I'm sure it's very necessary for presentation purposes but I'm not thinking that that will, in itself, address anything.

There's an example that you quote of Stuart Russell positing the family robot that cooks the cat for dinner because it seems like a good meal and the sort of mistake that an AI could make it but we're smart enough to do that, but not smart enough to know what the family valued in its pets. It seems to me that though that intersection, that gap of being smart enough to do one thing, but not the other might be empty. I have a hard time imagining that the prerequisites for a robot that has that amount of embodiment and general intelligence that it can conceive of the cat as being a source of protein and figure out how to cook it, is not going to understand the idea of using only ingredients that are in the pantry and following recipes that it can find or adapt. It seems that we should worry about that robot doing some things, but the cat ought to be safe. And I wonder if that example was chosen because it's got some instant meme appeal,

but whether there's a better one? I mean, I could probably get GPT-3 to say as much about how much humans value cats.

Maybe a family in Korea with the family dog would be a better example, but I definitely don't see how it's crazy. I mean, it's just the other white meat, right? You have a lot of human common sense so to you it seems like it's the wrong decision, but from point of view of cooking animals and nutritional value, the difference between a pig and a cat is completely arbitrary. We could have had pigs and farms for cats.

Right. It seems that we are seeing the boundaries of what we call common sense eroded by AI right now. I mean, some of the things that GPT-3 is saying would have been considered impossible to come out of the mouth of an AI a decade ago, I think. And we know how it's doing it. We know that there's no magic golem inside thinking of these things, but nevertheless, those results are surprising for what you can do with that size of a network. Could that line of development start demonstrating something that we would think of as common sense? Certainly, some of the things that I've seen it say, amount to common sense whether it had it internally or not. I'm trying to avoid sounding like John Searle and the Chinese room argument here, but it seems we're going to end up there.

So we can start by just agreeing on how difficult common sense actually is. It's very local, very culture-dependent, then you travel a lot, you realize that your common assertions are not at all common. Our people may not agree with what is for dinner as common animal, so it's extra difficult for AIs to align with us given that we don't agree, we don't align. If they do release GPT-4, GPT-5, I wouldn't be surprised if it got much better, it downloaded all the knowledge we ever published and there [are] enough examples of meals not being cats to where you can assume that's statistically unlikely, certainly.

Now, the control problem is directly related to the value alignment problem of aligning AI with our values. We have that problem right now with software. Anyone that's writing a program wants to align its results with their requirements and yet every software developer knows that the customer doesn't know what they want. They will tell you what they want, and you build that. And then not only are they going to say, "That's not what I meant", but they will also say, "That is what I meant but now I realize that was a mistake." And that's just with software right now. So as you point out in the paper, we don't even know what our own values are. How are we supposed to align AI with those? Do you have any thoughts about how we could get better at that?

No, but I do describe different levels of difficulty. So A) we don't know what we want. We also tend to change what we want as we get older or our position in society changes. If we knew what we want, and it was static - fixed - we wouldn't know how to code it in. We don't have programming language keyboards for many concepts like happiness or things of that nature. And even if somehow we magically managed to get that encoded in the machines, again, if they are smarter than us, then perhaps they'll advise us, "No, that is not what you want to want. You should want something else," like we do with children a lot of times. My children want me to

value align on their preference for ice cream. I try to give them healthy nutrients and we don't have alignment. I'm relatively superintelligent compared to them, and I do feel I'm a friendly parent, but they don't seem to be getting happier when I'm around.

I have exactly the same situation. There is a range of possible outcomes of developing an artificial superintelligence. And I realize that we keep saying "an" artificial superintelligence but the reality would surely be many of them not just instances of the same image, but different ones that would likely compete with each other. Can you describe what some of those are because they are on a spectrum from catastrophic to merely depressing?

So there are arguments that we're only going to get one - the first one we'll get will make sure that there is no competition. If it's sufficiently capable, it will prevent any others like with nuclear weapons - the countries who got there try to prevent other countries from getting to the same point, not only successfully but they do. As to what a possible consequence is you have full range from complete bliss to suffering risks and in the middle you have existential crisis where we are no longer around, so it's at zero. We're neither happy nor unhappy - we don't exist, but anything else, from positive infinity to negative infinity.

What's the complete bliss scenario?

So I have a paper where I describe scenarios where we are living in virtual universes and they are individual so every agent gets exactly what you want. It doesn't matter what you want. In your universe you can be king, you can be slave, it's up to you, and you're limited by your imagination. Now, this is to address the differences in what we want. So we don't have to agree, we don't have to have a democratic vote and force 49% to agree to 51%. You basically get exactly what you want and you can change that. You can travel to different universes experiencing them. All we have to do is control the substrate. If we manage to control the superintelligent substrate, you get your multiverse of human beings and not just human beings experiencing whatever they want to experience.

And I think at this point, not just the philosophers, but the psychologists are jumping up and down and saying, "That's not how people are supposed to live." It's clearly got some downside in terms of people that value autonomy and living in the real world as opposed to a simulation although there's also the argument that we're already living in a simulation.

Yes, I was about to say it really helps if you already think we are in a simulation. It definitely makes it easier to swallow.

Right. And my favorite solution to the Fermi paradox is that we're actually living in a simulation where the owner was too cheap to pay for the extragalactic species upgrade. So the more I read about the scenarios of artificial superintelligence and the ways in which they might try and do the right thing or help us, but that would be frustrating for us because we're still outclassed, the more I realized that Isaac Asimov had already thought about a number of these scenarios, just didn't have the same lexicon that we do now, and written them into his stories in some depth. Which actually made me think that there is one scenario I haven't seen listed, which is

that the artificial superintelligence decides that we should have our right to self-determination and it goes so long and leaves us to take care of ourselves and goes away. Which that certainly doesn't help if there are any other artificial superintelligences that don't share that but it seems worth adding to the list.

I would guess we would immediately try to build another one which is staying around and not abandoning us and provides helpful answers.

Right. Okay, talking about providing helpful answers, Douglas Adams went there. And I think that that had more education in it than we tend to recognize because it's very humorous, but he had Deep Thought, the computer that was asked the question, "What's the meaning of life, the universe and everything?" and took some long time to come up with the answer "42", and the joke was, "Well, we didn't understand the question properly." If we can use artificial superintelligence to build an oracle and we ask it those kinds of questions, those questions, that one, in particular, is not well-formed, can the answers it gives us even be useful?

It depends on the specific question. So, a question such as purpose of life may be very agent-dependent - you may not share purpose with others. So, a single answer for everyone would not be meaningful.

One of the things that we talk about in advanced AI is artificial general intelligence is understanding. And the word understand is one that the more I think about this, the less I actually understand what that word means. And this is getting back again to Searle and the Chinese room where he argues that the agent in the room or the room itself does not understand Chinese. I'll explain that one in the show opening. But do you ascribe to artificial general intelligence, the quality of understanding or is that one of those words that we should toss out as not being useful?

No, I think it's important that it probably has something to do with AI completeness - ability to solve any problem a human can solve, and to be able to solve certain problems, you need to create this nearly complete model of the universe, of a domain you're working in. And I think we can call that understanding where you have a model which corresponds maybe not a hundred percent accurately to what you're doing, but to a high degree.

Okay, so I hope no one was too disturbed by all the talk about cooking cats – no cats were harmed in the making of this interview – but you'll appreciate that we're trying to head off a future where AI makes that kind of mistake and the family pet becomes the family dinner.

In today's look at the AI headlines, the Defense Advanced Research Projects Agency – or DARPA, America's favorite mad scientist think-tank, just hosted what it called the AlphaDogfight challenge. Now you may that DARPA hosted the *Grand Challenge* in the early 2000s to find a winner in the autonomous vehicle competition. And at that time, the best entrant could go manye a mile without falling off the road. And we know what we have now. Well, what's DARPA doing now? DARPA is the government agency responsible for the development of emerging technologies. And they invited companies to build their own AI to compete in a fighter pilot simulation.

And in August they selected eight companies to compete in the final challenge. The list included Lockheed Martin as well as several smaller companies. Now the AIs that these companies built were pitted against each other as well as a top human F-16 fighter pilot in the United States Air Force. And the result was conclusive. The winner was an AI from a company called Heron Systems, which beat all seven of the AI competitors and absolutely destroyed the human fighter pilot, winning five rounds to zero.

A big component of its success was its ability to adapt on the fly. The human fighter pilot said that by round five, all the normal tactics he employed weren't working against the AI. It anticipated and countered his every move. This is very much reminding me of a *Star Trek: The Next Generation* episode called the *Arsenal of Freedom*. So the human pilot was forced to shift his tactics, but he couldn't adapt fast enough. Is this the shape of combat to come? Why wouldn't it be?

In the last episode with Karina Vold we touched on the King Midas problem, which is also known as the genie problem, and here we got into it more detail. Even though we don't know how far off the scenario is of artificial superintelligences – or even just one – possessing immense power, the question of how we align its values with ours is so difficult it's a good thing that people like Roman are tackling it right now. Philosophers have spent thousands of years trying to unpack what the phrase "our values" means and here we are maybe needing to understand it extremely well within the next fifty years. We'll open that up even more in the next episode when we conclude the interview with Roman Yampolskiy.

Until then, remember: no matter how much computers learn how to do, it's how we come together as humans that matters.

http://aiandyou.net