AI and You

Transcript

Guest: Roman Yampolskiy

Episode 17

First Aired: Monday, October 12, 2020

Hi, welcome to episode 17. Today we're going to conclude the interview with Dr. Roman Yampolskiy. He is a tenured associate professor in the department of Computer Engineering and Computer Science at the University of Louisville in Kentucky where he founded and directs the Cyber Security lab. Some of those include that he is also a senior member of the IEEE and a Research Advisor for the Machine Intelligence Research 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. He has made numerous media appearances and conference presentations.

In the last episode we talked mostly about his new paper, titled "On Controllability of Artificial Intelligence." Roman described the Control Problem of AI in that episode, which is something people concerned with the future of AI talk about a lot, and it boils down to, how do we ensure that AI remains under our control when it becomes much more intelligent than us? That intelligence would lead it to become more powerful, and that leads us to the value alignment problem – how do we align the values of that very smart AI with our own? This is a problem when we don't even know what human values are. They change from culture to culture, year to year, person to person, situation to situation. Psychologists love to conduct studies that demonstrate how inconsistent we are.

So the Control Problem is like having a genie, only you have more than three wishes. But the genie takes your instructions very literally and if you are wrong about how it will interpret them you may not get the "undo" wish. Tell it you want to end the climate crisis and it may decide to wipe out humans because we are the biggest cause of the climate crisis. Oops.

When it comes to creating perfect instructions, we suck. Courts are kept constantly busy by the imperfection of our laws. There was a practice that was a thing when I was a kid in Britain, called the Work to Rule. It was a tactic of unions when they wanted bargaining power. It meant they would instruct their members to do everything exactly as specified by the company policy and procedures manual. If the manual said they had to get manager approval before touching a lathe they hadn't used before, then they wouldn't touch that lathe without say-so, even if the manager wasn't in, even if getting the product out the door depended on it. Whereas normally they would bend the rule for the sake of getting the job done. We're all familiar with doing that, right? Does there exist any workplace whose procedures are completely fit for all tasks and kept constantly up to date? So of course the management could only fume and go back to the bargaining table, but they couldn't take any action against the workers because they were doing *exactly* what the management had said.

There are some people who think that the only problem here was that the manual wasn't accurate enough. They think that you could have a perfect manual. These are probably also the people who think there would be no trouble telling a superintelligence what to do. Let's hear what Roman thinks about that and what we should do about superintelligent AI in the second half of the interview.

I want to make sure that I raise something here that is perhaps frowned on in academic circles, I don't know, but that's the issue of the question of feelings in an artificial intelligence. It was actually only relatively recently that I realized that the word sentience doesn't mean "thinking" it actually means "feeling" and had been used all the time in science fiction talking about sentient races, but sentience actually means "feeling", and it's sapience that means "thinking" and that's why we're *Homo Sapiens*. Now, these assumptions about AIs, artificial superintelligences seem to be that they don't have feelings, they don't have emotions, and that the problem for us arises from the fact that we do, otherwise we might be quite content to be outclassed by artificial superintelligence. This raises a number of questions, but one is, what is the likelihood that developing an artificial superintelligence or even on artificial general intelligence, that it would not necessarily come along with feelings?

So this sort of thing usually I talk about consciousness and qualia, not so much directly feelings but ability to experience something. What is it like to feel pain, pleasure? And I do have a paper where I argue that it's a side effect of computation, kind of like heat being released as a side effect of computation, having those experiences which are based on your hardware and algorithm you're running. So I talk about optical illusions and how you experience them. So sometimes you get a really good one and you see things rotating and colors changing, and maybe someone else with different optical system doesn't, but some artificial neural networks do. And so I argued that they have rudimentary states of consciousness - they experience those illusions. So there is a chance that you'll get consciousness for free as a side effect of developing intelligence and it's proportionate to how intelligent, how complex it is. So superintelligence would be super conscious. It would have deeper feelings. It would experience pain and pleasure in ways we can never comprehend.

And the terms you're using there remind me that there are a number of things that arrive all at the same time in human beings. We have intelligence, we have consciousness, we have self-determination, free will, independent thought, we have creativity, we have feelings, and yet there's perhaps no guarantee that those things all arrive along with a survival instinct. In an artificial general intelligence, they could be distinct. They perhaps go together in humans because there was some evolutionary imperative for it like as we evolved consciousness, there was a reason for also evolving emotions and feelings to go along with that. And maybe we'll discover that there's a similar imperative as we create artificial general intelligence. I don't know, but to what extent can those things be separated so that you could get some without the others in an artificial general intelligence?

I think it might be possible, especially if you're trying to do it on purpose. The reason I think that [is] some human beings don't experience certain qualia. Some, for example, don't see certain things others do, some cannot percept faces. I kind of like reading through those things on Wikipedia. There are hundreds of disability conditions which prevent you from experiencing certain qualia. Color blindness is an obvious example but it seems like it's possible to be human-level intelligent and not have certain types of conscious experiences. So probably AIs will be likewise capable of separating optimization from experience of that type.

And we tend to think that an artificial superintelligence would be a human brain but with all of the power of computers behind it and computer processing power, it would be our intelligence on steroids. But if I think about what it would be like for a human to have all of Google inside their brain, the brain would probably explode. I wonder whether we might find limitations to what this level, our level of general intelligence can sustain in terms of that amount of information processing power. Does that make any sense?

Absolutely. In fact, I have a recent paper saying that humans are not AGIs. We are human-level, we're not general in a true sense. We are general in the human domain of interest and expertise but there are things we can never do which a general AI would be able to do. And I also often take on this idea that we can merge with machines, where we would be part of them and work side by side somehow integrated superintelligence in the human brain, but I never understood what it is we're offering in that equation. We have limited memory. We cannot comprehend the big picture. At best, we'd become some sort of ignored component and bottleneck and eventually, we're completely removed or bypassed. So I never understood how that solution is supposed to work.

There's a naughty part of me that wants to ask whether this is the perfect field for you because it's the ultimate expression of Russian fatalism.

This is above my pay grade.

You talk in the paper about losing control. The whole control problem is about losing control of artificial intelligence and yet I wonder right now as a species, just how much we actually do control. I mean, arguably looking at the headlines, there's a lot of things we ought to be controlling that we aren't. Talk about what sort of control you expect or foresee us losing in these scenarios.

So right now, I think we kind of willingly give up a lot of control, [whether] it's through overindulging in social media or letting computers run the stock market. But if we decided, we can stop. We can say, "Okay, no more bots trading stocks. Has to be a human." We are making that decision. We're still in charge. I think some of my concerns now that we get to a point of no return where you can no longer unplug the system, there is no undo bottom, someone else, not a human is deciding the future and we are not impacting that decision.

It's analogous to, as a thought experiment, the earth being visited by aliens which would be much more capable than us in order to be getting here before we got there, and superintelligent and so we would be overshadowed. Living in Canada at the moment, I'm aware of what it's like to be overshadowed by a neighbor, and there are various, both polite and unkind metaphors for that comparison. But nevertheless, Canada is still here; Canadians are still here. Maybe the US isn't that much bigger than Canada than an artificial superintelligence would be but is it perhaps a uniquely American phenomenon to be worried about being overshadowed by something when you're arguably the top dog in so many ways?

I think it's a human condition. We are probably not the best at anything individually. I'm not the fastest human. I'm not the smartest human so I understand very well what it's like to compete with those who are better than me but what if there was one agent which was better than me at everything? That would not be a very fair competition. I definitely would not win. And now it's one thing to simply lose a competition it's a different thing to lose everything else based on who wins this competition. If there is a possibility of that happening, it is a concern. I've seen people argue that it's only white males who worry about such problem or only Americans who worry about such problem but I think anyone who does work in AI safety, whether they are in Canada, or they may be not a male, will have you similar concerns.

We were talking earlier about Eliezer Yudkowsky and you quote him as saying both "We can't know when this is going to happen" and also placing a bet that the world would be destroyed by out of control artificial intelligence by January 1, 2030. And note that he's doing that in order to bring attention to the issue because obviously, it wouldn't be in a position to collect on that. But that does put a stake in the ground as far as a timeline and pretty aggressive one by most standards. Is there work that you would like to see being done on this problem right now that isn't?

Well, not maybe a specific project, but just in general, the amount of resources allocated to it. I think there [are] maybe a dozen people in the world who are dedicated full-time to superintelligence and maybe [the] majority of them are in philosophy departments. So it seems like given how much we put into cures for baldness and impotence and things of that nature, this may maybe get 1% of those resources both in terms of resources of mental dedication and financial resources. It seems like it might be useful to know such things.

A very logical argument but I'm continually reminded of how we don't collectively make decisions based on logic so much as emotion. You mentioned climate change earlier, and that's something where there were projections over 100 years ago, I read, of the carbon dioxide being put into the atmosphere causing the effects that we're seeing now. Even 40 years ago, if we'd acted on what was happening, we'd be in a different world, but now, we aren't. And yet, we still have the United States pulling out of the Paris climate accords and acting as though it's not a foregone conclusion. Climate change is both useful in the sense that we have some immediate effects, and then there are long term effects that are even more catastrophic so that it's analogous to the AI control problem in that way. But it's unhelpful in another sense in that with climate change, you can draw a line on a chart and say, "This is when the Maldives is underwater. This is when Miami floods. This is when Phoenix becomes uninhabitable" and we can't do that. So it makes it hard to get more people to take it seriously so that there are more than a dozen people working on that. Is there work that we could do that would amplify the immediacy, urgency, or importance of this?

Well, I think the field of AI itself is working on that. Every time there is a breakthrough, every time we get something like GPT-3, more and more people go “Oh, I didn't expect that. That surprised me. How else will I be surprised next year?” So I think that's the best we can hope for. Some people talk about maybe this is informational hazard. Some people talk about purposeful accidents, right? If you can use AI to cause some minor damage to prove a point, maybe that will help, but actually, it doesn't help. Such things are kind of like vaccination. They just go “See, nothing happened. We don't have to worry about it.” But yeah, to me, the progress in AI is the biggest indicator. It’s like watching water rise. You can see the cities slowly drowning. You can see capabilities go up every year.

It seems that the accidents are going to take care of themselves. If someone was writing Stuxnet today and they added AI to it, the potential for mayhem would be on a much larger scale because that experienced its own bug and infected many more computers than it was supposed to.

Right. So we definitely see examples of viruses acting not as they were programmed. There was a mistake and they became much more viral and impacted more machines. If they had capability for social engineering, they would certainly penetrate almost every network. Definitely something to worry about from the security point of view of that.

Can you share some of the other contributions that you've made? You coined the term intellectology, I believe?

So I'm looking at studies of intelligence and it seems like different fields I will interested in it, but they use different terms, whether you’re in philosophy, psychology, computer science. Dozens of fields look at how to create intelligence, how to measure it, how to develop it, how to detect it, how to detect artifacts produced by intelligence, but all of it is very separated. We don't have common terms. We don't have common tools. So one idea was to kind of bring it into its own discipline and share those resources. Hopefully, it will help our progress and you can see AI as one subarea within that field. How do we create different intelligences? Can they be in different substrates? What are the different types of minds?

I'm thinking about how human minds have various failure modes. There's a whole book, DSM-IV, of those and how different those are from the kind of failure modes that we currently experience in computers, particularly, for instance, image recognition right now. It looks on the surface as though recognition of say, stop signs on the road is happening the same way that we do it because we can find neurons that respond to diagonal lines and so forth. And yet those can be defeated by the sort of graffiti that no human would ever be fooled by, so they're not thinking the same way that we are. Will we need to develop different models to make more inroads on developing artificial general intelligence like developing associational memory models to make more of a parallel with the way that humans store and access memories?

Well, it's not obvious that we want systems to work the way human mind works. As you said, there are quite a few bugs. We can also talk about cognitive biases. Ideally, we want our intelligent systems to not have those. They may be a desirable feature. I have a paper about artificial stupidity where we on purpose introduced those to make systems more human-like, better products, easier to use. But in some cases, I think we want a system not to have bugs built on purpose, and not to have limitations of the human mind, so they better assist us. If I don't see something the system would see it. So creating identical clone may not be what we want in many cases.

I wasn't thinking so much of a clone like a mind upload but in this case, it seems that we may hit a wall in the application of deep learning and that we need other models to develop artificial general intelligence. What's your thinking about that?

So far, I see it progressing really well. I don't see any upper bound. We keep adding compute and big data and it keeps improving. We may need multiple models trained in different ways, but the general principle seems to be holding up.

When we train our models, they're starting from scratch. Human learns how to recognize an object after having spent years on the planet navigating around and mastering a whole lot of object recognition techniques that our AI models don't have. They're starting from beneath amoeba level every time. And so they need this huge amount of data just to be able to recognize a cat, whereas a human being can do it with one cat. There seems to be a fundamental difference there that if we're going to develop an artificial general intelligence that can do one-shot learning, we've got to have different substrates for doing that. What do you think?

Well, you said yourself that we spent years trying to build up this toolset of recognizing objects and all that so no, we don't recognize a cat right away. We spent years trying to figure out what kind of animal it is. Is it a plant? So I think in terms of amount of exposure to data, both artificial neural networks and human neural networks probably get about the same amount of data. We just kind of start testing it with a cat or a new animal. "Oh, look, is this *gof gof*, or what kind of animal is this?" But you forget all the years of pre-processing it took.

It seems to me that we need to develop more libraries. Like we have TensorFlow and so now it becomes possible to do image tagging and a few lines of code because we've got that library behind it. If we want embodied agents to understand the real world, which is what artificial general intelligence requires, it seems we need that kind of library to pass around about an understanding of the real world. Would that make sense because we don't have that or anything like that at the moment?

Right. It would be useful. We don't want to re-learn everything from scratch. Once we have a system which can reliably identify objects, for example, anyone else can use it. There are attempts to create databases like that of facts about life. So just hard coding, manually coding millions of pieces of information so other AI systems can use it and hopefully continue this process and learn them automatically from the internet, from Wikipedia. So yeah, eventually there is going to be a common database of existing AI capabilities, I'm sure.

Do you think we'll see another AI winter? Certainly, the hype cycle is driven by hype rising to the point where the technology can no longer fulfill the promises. Have we reached that level?

I would say unlikely given the amount of resources. We never had so much, not just from government funding agencies, but from corporations. Every company, every business has AI lab, AI research, products they need developed so it would be surprising if all of it just died out and we went back to hiring more humans to type things in. It's possible, I guess, but I would bet against it.

Well, the applications wouldn't be thrown out but there have been periods in the past where AI was a dirty word; you had to say you were working on something else. And I wonder whether we could be precipitated into another one of those when self-driving cars don't live up to the hype that's been promoted for them for the last few years.

I understand obviously, we're not going to just remove everything we already have, but I think in the past, let's say you had a government agency providing 90% of funding, if they decided to switch topics, you couldn't develop your expert systems anymore. Whereas right now, I think even if NSF decides not to fund the AI, which they just did for another hundred million this week, I think we'll still have enough interest from private sector, from other countries where-- We got to the point where it's a self-sustaining cycle. You need to keep developing it to stay in business so I don't think it's going to stop for a long time. It may slow down in that particular direction if we hit some bottlenecks, but there [are] a lot of alternative paths we can explore, and, again, I think the resources are there for it.

So [to] wrap up here, there are so many more questions, but hopefully, we can leave those for another time. Ten years from now, let's say Eliezer lost the bet, what would you like to be true? What do you think will be true in artificial intelligence then?

So, predicting [the] future is hard, especially for very specific time horizons. I am much more comfortable talking about general trends towards the future. I can tell you this is more likely to be so, again, like with AI winter I think it's less likely. Specifically what happens in 10 years is just such a hard question. I think in many ways, we'll be beyond where we think we're going to be. So I think self-driving will be totally done - not a problem at all, maybe even sooner. Whereas it’s possible many other things will turn out to be much harder. But I definitely don't think I have the magical power of predicting [the] future that precisely.

Done, but point taken. I think was Bill Gates who said, “We tend to overestimate what we can do in two years and underestimate what we can do in ten” because we don't have intuitive grasp - well, he didn't add this, but I did, that we don't have an intuitive grasp of exponential growth.

We get surprised. Something which we didn't know is likely to work happens to work really well. Somebody comes up with a new model, new paper, and it changes everything in a year or two.

And so what's next for you? Any new papers, books in the pipeline?

Yes, I'll continue my set of papers and limits to what we can do in the space. I now have a paper on “Unpredictability of AI”, unexplainability, incomprehensibility and controllability. I suspect you'll see a few more trying to make similar points, complete the set. I think it's very important to know what we can do so we know where resources are to go, and obviously, you need to know what you can’t do no matter how much you get in terms of resources. It's standard computer science. We know certain problems are unsolvable and we need to find workarounds to get to the good place without trying to solve impossible problems.

Thank you. Where should people go to find out more about what you're doing, get your books, learn more about your work?

You can follow me on Facebook. You can follow me on Twitter. Please don't follow me home.

You've heard that question before.

Thank you, Roman. It's been an extreme pleasure having you on here. You really range over some wide horizons of time and intellectual thought here.

Thank you so much for inviting me. I hope to be back.

There you go. What do you think, should we ask him back? I think there’re a lot more questions we could ask Roman. If you want to send them to me, I’ll make a collection.

We talked about how hard it is to predict the future. The introduction of Deep Learning around 2010 advanced the progress of artificial intelligence to a degree previously unimaginable. The accuracy of image and speech recognition, for instance, that deep learning made available, were just pipe dreams before. The history of the Google Translate engine is a good example. When they put deep learning on the task of translating between human languages, the jump in performance was so great they didn’t even announce it, just waited to see who would notice, and people were talking about it within hours. We can extrapolate the acceleration of our computer hardware fairly well, but what new breakthrough in our software, like deep learning, lies just around the corner, waiting to bring those artificial superintelligences a step closer? Time to start polishing the genie lamp.

Next week’s guest will be Tony Czarnecki, an economist and futurist calling in from the United Kingdom to talk about his latest book, “Becoming a Butterfly.” What’s it about? Well, it’s the latest in his Posthuman series, and it’s about… transformation. That’s next week on Artificial Intelligence and You.

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