Lex Fridman: What to you is the probability that superintelligent AI will destroy all human civilization?  
  
Roman Yampolskiy: What's the timeframe?  
  
Lex Fridman: Let's say a hundred years, in the next hundred years.  
  
Roman Yampolskiy: So the problem of controlling AGI or superintelligence, in my opinion, is like the problem of creating a perpetual safety machine by analogy with a perpetual motion machine. It's impossible. Yeah, we may succeed and do a good job with GPT-5, 6, 7, but they just keep improving, learning, eventually self-modifying, interacting with the environment, interacting with malevolent actors. The difference between cybersecurity, narrow AI safety, and safety for general AI or superintelligence is that we don't get a second chance with cybersecurity. Somebody hacks your account; what's the big deal? You get a new password, a new credit card, you move on. Here, if we're talking about existential risks, you only get one chance. So you're really asking me what are the chances that we'll create the most complex software ever on the first try with zero bugs? And it will continue to have zero bugs for 100 years or more.  
  
Lex Fridman: So there is an incremental improvement of systems leading up to AGI. To you, it doesn't matter if we can keep those safe. There's going to be one level of system at which you cannot possibly control it.  
  
Roman Yampolskiy: I don't think we have so far made any system safe at the level of capability they display. They already have made mistakes. We had accidents; they've been jailbroken. I don't think there is a single large language model today that no one has been successful at making do something developers didn't intend to do.  
  
Lex Fridman: But there's a difference between getting it to do something unintended, getting it to do something that's painful, costly, destructive, and something that's destructive to the level of hurting billions of people, or hundreds of millions of people, or the entirety of human civilization. That's a big leap.  
  
Roman Yampolskiy: Exactly. But the systems we have today have the capability of causing X amount of damage. So then they fail. That's all we get. If we develop systems capable of impacting all of humanity or the entire universe, the damage is proportionate.  
  
Lex Fridman: What do you see as the possible ways that such a kind of mass murder of humans can happen?  
  
Roman Yampolskiy: It's always a wonderful question. So one of the chapters in my new book is about unpredictability. I argue that we cannot predict what a smarter system will do. So you're really not asking me how superintelligence will kill everyone; you're asking me how I would do it. And I think it's not that interesting. I can tell you about the standard nanotech, synthetic biology, nuclear superintelligence, but it will come up with something completely new, completely super. We may not even recognize that as a possible path to achieve that goal.  
  
Lex Fridman: So there is like an unlimited level of creativity in terms of how humans could be killed. But, you know, we could still investigate possible ways of doing it, not how to do it. But at the end, what is the methodology that does it? You know, shutting off the power and then humans start killing each other, maybe because the resources are really constrained. Then there's the actual use of weapons, like nuclear weapons, or developing artificial pathogens, viruses, that kind of stuff. We could still kind of think through that and defend against it, right? There's a ceiling to the creativity of mass murder of humans here, right? The options are limited.  
  
Roman Yampolskiy: They are limited by how imaginative we are. If you are that much smarter, that much more creative, if you are capable of thinking across multiple domains, doing novel research in physics and biology, you may not be limited by those tools. If squirrels were planning to kill humans, they would have a set of possible ways of doing it, but they would never consider things we can come up with.  
  
Lex Fridman: So are you thinking about mass murder and destruction of human civilization? Are you thinking of... with squirrels, you put them in a zoo and they don't really know they're in a zoo. If we just look at the entire set of undesirable trajectories, the majority of them are not going to be death. Most of them are going to be just like things like "Brave New World," where, you know, the squirrels are fed dopamine and they're all doing some kind of fun activity, and the soul of humanity is lost because of the drug that's fed to it. Or like literally in a zoo. We're in a zoo. We're doing our thing. We're like playing a game of Sims, and the actual players playing that game are AI systems. Those are all undesirable because the fire of human consciousness is dimmed through that process, but it's not killing humans. So, us thinking about that, or is the biggest concern literally the extinction of humans?  
  
Roman Yampolskiy: I think about a lot of things. So there is existential risk, where everyone's dead. There is suffering risk, where everyone wishes they were dead. We also have the idea of Ikigai risk, where we lost our meaning. The systems can be more creative; they can do all the jobs. It's not obvious what you have to contribute to a world where superintelligence exists. Of course, you can have all the variants you mentioned where we are safe, we are kept alive, but we are not in control. We are not deciding anything. We are like animals in a zoo. There are possibilities we can come up with as very smart humans, and then possibilities something a thousand times smarter can come up with for reasons we cannot comprehend.  
  
Lex Fridman: I would love to sort of dig into each of those X risk, S risk, and Ikigai risks. So can you linger on Ikigai? What is that?  
  
Roman Yampolskiy: So the Japanese concept of Ikigai is finding something that allows you to make money, you are good at it, and society says we need it. So, like, you have this awesome job; you are a podcaster, it gives you a lot of meaning, you have a good life, I assume you're happy. That's what we want most people to find. For many intellectuals, it is their occupation which gives them a lot of meaning. I am a researcher, philosopher, scholar. That means something to me. In a world where an artist is not feeling appreciated because his art is just not competitive with what is produced by machines, or a writer or scientist will lose a lot of that. And at the lower level, we're talking about complete technological unemployment. We're not losing 10% of jobs; we're losing all jobs. What do people do with all that free time? What happens then? Everything society is built on is completely modified in one generation. It's not a slow process where we get to figure out how to live that new lifestyle, but it's pretty quick.  
  
Lex Fridman: In that world, can't humans do what humans currently do with chess? Play each other, have tournaments? Even though AI systems are far superior this time in chess. So we just create artificial games, or for us, they're real, like the Olympics. And we do all kinds of different competitions and have fun, focusing on maximizing the fun, and let the AI focus on the productivity.  
  
Roman Yampolskiy: It's an option. I have a paper where I try to solve the value alignment problem for multiple agents. And the solution to avoid compromise is to give everyone a personal virtual universe. You can do whatever you want in that world. You could be a king, you could be a slave; you decide what happens. So it's basically a glorified video game where you get to enjoy yourself and someone else takes care of your needs. And the substrate alignment is the only thing we need to solve. We don't have to get 8 billion humans to agree on anything.  
  
Lex Fridman: Okay, so why is that not a likely outcome? Why can't AI systems create video games for us to lose ourselves in, each with an individual video game universe?  
  
Roman Yampolskiy: Some people say that's what happened.  
  
Lex Fridman: We're in a simulation and we're playing that video game. And now we're creating what maybe we're creating artificial threats for ourselves to be scared about, because fear is really exciting. It allows us to play the video game more vigorously.  
  
Roman Yampolskiy: And some people choose to play on a more difficult level with more constraints. Some say, okay, I'm just going to enjoy the game. High privilege level. Absolutely.  
  
Lex Fridman: So, okay, what was that paper on multi-agent value alignment?  
  
Roman Yampolskiy: Personal universes. Personal universes.  
  
Lex Fridman: So that's one of the possible outcomes. But what in general is the idea of the paper? It's looking at multiple agents that are human-AI, like a hybrid system where there's humans and AIs, or is it looking at humans or just intelligent agents?  
  
Roman Yampolskiy: In order to solve the value alignment problem, I'm trying to formalize it a little better. Usually, we're talking about getting AIs to do what we want, which is not well defined. Are we talking about the creator of a system, the owner of that AI, humanity as a whole? But we don't agree on much. There is no universally accepted ethics or morals across cultures, religions. People have very different preferences politically and such. So even if we somehow managed all the other aspects of it, programming those fuzzy concepts and getting AIs to follow them closely, we don't agree on what to program in. So my solution was, okay, we don't have to compromise; you have your universe, I have mine, whatever you want. And if you like me, you can invite me to visit your universe. We don't have to be independent, but the point is you can be. And virtual reality is getting pretty good. It's going to hit a point where you can't tell the difference. And if you can't tell if it's real or not, what's the difference?  
  
Lex Fridman: So basically, give up on value alignment, create an entire... it's like the multiverse theory. Just create an entire universe for you. What your values, you still have to align with that individual. They have to be happy in that simulation. But it's a much easier problem to align with one agent versus 8 billion agents plus animals, aliens.  
  
Roman Yampolskiy: I'm trying to do that. Yeah.  
  
Lex Fridman: Okay, is there any way to... So, okay, that's giving up on the value alignment problem. Well, is there any way to solve the value alignment problem where there are a bunch of humans, multiple humans, tens of humans, or 8 billion humans that have a very different set of values?  
  
Roman Yampolskiy: It seems contradictory. I haven't seen anyone explain what it means outside of words, which pack a lot. Make it good, make it desirable, make it something they don't regret. But how do you specifically formalize those notions? How do you program them in? I haven't seen anyone make progress on that so far.  
  
Lex Fridman: But isn't that the whole optimization journey that we're doing as a human civilization? We're looking at geopolitics. Nations are in a state of anarchy with each other. They start wars; there's conflict, and oftentimes they have very different views of what is good and what is evil. Isn't that what we're trying to figure out, just together trying to converge towards that? So we're essentially trying to solve the value alignment problem with humans.  
  
Roman Yampolskiy: Right, but the examples you gave, some of them are, for example, two different religions saying this is our holy site and we are not willing to compromise it in any way. If you can make two holy sites in virtual worlds, you solve the problem, but if you only have one, it's not divisible. You're kind of stuck there.  
  
Lex Fridman: But what if we want to be in tension with each other and that through that tension we understand ourselves and we understand the world? So that's the intellectual journey we're on. We're on as a human civilization. We create intellectual and physical conflict and through that figure stuff out.  
  
Roman Yampolskiy: If we go back to that idea of simulation and this is entertainment giving meaning to us, the question is how much suffering is reasonable for a video game. So yeah, I don't mind a video game where I get haptic feedback that is a little bit of shaking. Maybe I'm a little scared. I don't want a game where kids are tortured, literally. That seems unethical, at least by our human standards.  
  
Lex Fridman: Are you suggesting it's possible to remove suffering? If we're looking at human civilization as an optimization problem.  
  
Roman Yampolskiy: So we know there are some humans who, because of a mutation, don't experience physical pain. So at least physical pain can be mutated out, reengineered out. Suffering in terms of meaning, like you burn the only copy of my book, is a little harder. But even there you can manipulate your hedonic set point; you can change defaults, you can reset. The problem with that is if you start messing with your reward channel, you start wireheading and end up blissing out a little too much.  
  
Lex Fridman: Well, that's the question. Would you really want to live in a world where there's no suffering? That's a dark question. Is there some level of suffering that reminds us of what this is all for?  
  
Roman Yampolskiy: I think we need that. But I would change the overall range. So right now it's negative infinity to kind of positive infinity—pain and pleasure access. I would make it like zero to positive infinity. And being unhappy is like I'm close to zero.  
  
Lex Fridman: Okay, so what's the S risk? What are the possible things that you're imagining with S risk? So mass suffering of humans, what are we talking about there? Caused by AGI?  
  
Roman Yampolskiy: So there are many malevolent actors we can talk about—psychopaths, crazies, hackers, doomsday cults. We know from history they tried killing everyone; they tried on purpose to cause the maximum amount of damage, terrorism. What if someone malevolent wants to torture all humans for as long as possible? You solve aging, so now you have functional immortality, and you just try to be as creative as you can.  
  
Lex Fridman: Do you think there are actually people in human history who tried to literally maximize human suffering? Just studying people who have done evil in the world, it seems that they think that they're doing good. It doesn't seem like they're trying to maximize suffering; they just cause a lot of suffering as a side effect of doing what they think is good.  
  
Roman Yampolskiy: So there are different malevolent agents; some may just gain personal benefit and sacrifice others to that cause. Others we know for effect try to kill as many people as possible. When we look at recent school shootings, if they had more capable weapons, they would take out not dozens, but thousands, millions, billions.  
  
Lex Fridman: Well, we don't know that, but that is a terrifying possibility. And we don't want to find out. Like if terrorists had access to nuclear weapons, how far would they go? Is there a limit to what they're willing to do? In your sense, are there some malevolent actors where there's no limit?  
  
Roman Yampolskiy: There are mental diseases where people don't have empathy, don't have this human quality of understanding suffering in others.  
  
Lex Fridman: And then there's also a set of beliefs where you think you're doing good by killing a lot of humans.  
  
Roman Yampolskiy: Again, I would like to assume that normal people never think like that. It's always some sort of psychopaths.  
  
Lex Fridman: But yeah, and to you, AGI systems can carry that and be more competent at executing that.  
  
Roman Yampolskiy: They can certainly be more creative. They can understand human biology better, understand our molecular structure, our genome. Again, a lot of times torture ends when an individual dies. That limit can be removed as well.  
  
Lex Fridman: So if we're actually looking at X risk and S risk as the systems get more and more intelligent, don't you think it's possible to anticipate the ways they can do it and defend against it, like we do with cybersecurity? We do security systems, right?  
  
Roman Yampolskiy: We can definitely keep up for a while. I'm saying you cannot do it indefinitely. At some point, the cognitive gap is too big. The surface you have to defend is infinite. But attackers only need to find one exploit.  
  
Lex Fridman: So to you, eventually, this is—we're heading off a cliff.  
  
Roman Yampolskiy: If we create general superintelligences, I don't see a good long-term outcome for humanity. The only way to win this game is not to play it.  
  
Lex Fridman: Okay, well, we'll talk about possible solutions and what not playing it means. But what are the possible timelines here to you? What are we talking about? Are we talking about a set of years, decades, centuries? What do you think?  
  
Roman Yampolskiy: I don't know for sure. The prediction markets right now are saying 2026 for AGI. I heard the same thing from the CEO of Anthropic and DeepMind. So maybe we are two years away, which seems very soon given we don't have a working safety mechanism in place or even a prototype for one. And there are people trying to accelerate those timelines because they feel we're not getting there quickly enough.  
  
Lex Fridman: But what do you think they mean when they say AGI?  
  
Roman Yampolskiy: So the definitions we used to have, and people are modifying them a little bit lately, Artificial General Intelligence was a system capable of performing in any domain a human could perform. So kind of you are recreating this average artificial person; they can do cognitive labor, physical labor, where you can get another human to do it. Superintelligence was defined as a system that is superior to all humans in all domains. Now people are starting to refer to AGI as if it's superintelligence. I made a post recently where I argued that for me, at least, if you average out over all the common human tasks, those systems are already smarter than the average human. So under that definition, we have it. ChainLA has this definition of where you're trying to win in all domains. That's what intelligence is. Now, are they smarter than elite individuals in certain domains? Of course not. They're not there yet. But the progress is exponential.  
  
Lex Fridman: See, I'm much more concerned about social engineering. So to me, AI's ability to do something in the physical world is like the lowest hanging fruit. The easiest set of methods is by just getting humans to do it. It's going to be much harder to be the kind of virus that takes over the minds of robots, where the robots are executing the commands. It just seems like humans social engineering of humans is much more likely.  
  
Roman Yampolskiy: That would be enough to bootstrap the whole process.  
  
Lex Fridman: Okay, just to linger on the term AGI. What to you is the difference between AGI and human-level intelligence?  
  
Roman Yampolskiy: Human level is general. In the domain of expertise of humans, we know how to do human things. I don't speak dog language. I should be able to pick it up. If I'm a general intelligence, it's kind of inferior to an animal. I should be able to learn that skill, but I can't. A general intelligence, truly universal general intelligence should be able to do things like that humans cannot do.  
  
Lex Fridman: To be able to talk to animals.  
  
Roman Yampolskiy: For example, to solve pattern recognition problems of that type, to do similar things outside of our domain of expertise, because it's just not the world we live in.  
  
Lex Fridman: If we just look at the space of cognitive abilities we have, I just would love to understand what the limits are beyond which an AGI system can reach. What does that look like? What about actual mathematical thinking or scientific innovation?