Lex Fridman Podcast #431 - Roman Yampolskiy: Dangers of Superintelligent AI

Published - June 2, 2024

Transcribed by - thepodtranscripts.com

Lex Fridman
The following is a conversation with Roman Yampolskiy, an AI safety and security researcher and author of a new book titled AI: Unexplainable, Unpredictable, Uncontrollable. He argues that there's almost 100% chance that AGI will eventually destroy human civilization. As an aside, let me say that I'll have many often technical conversations on the topic of AI, often with engineers building the state-of-the-art AI systems. I would say those folks put the infamous P(doom) or the probability of AGI killing all humans at around 1 to 20%, but it's also important to talk to folks who put that value at 70, 80, 90, and is in the case of Roman, at 99.99 and many more nines percent. I'm personally excited for the future and believe it will be a good one in part because of the amazing technological innovation we humans create, but we must absolutely not do so with blinders on ignoring the possible risks, including existential risks of those technologies. That's what this conversation is about. This is the Lex Fridman podcast. To support it, please check out our sponsors in the description. Now dear friends, here's Roman Yampolskiy. What to you is the probability that super intelligent 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 a problem of creating a perpetual safety machine. By analogy with perpetual motion machine, it's impossible. Yeah, we may succeed and do 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 for 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, new credit card, you move on. Here, if we're talking about existential risks, you only get one chance. So you are 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'll continue to have zero bugs for a hundred 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 so far have 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, which no one was successful at making do something developers didn't intend it to do.

Lex Fridman
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, billions of people, or the entirety of human civilization. That's a big leap.

Roman Yampolskiy
Exactly, but the systems we have today have capability of causing X amount of damage. So when we fail, that's all we get. If we develop systems capable of impacting all of humanity, all of universe, the damage is proportionate.

Lex Fridman
What to you are the possible ways that such 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. I think it's not that interesting. I can tell you about the standard nanotech, synthetic, bio, nuclear. Superintelligence 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 an unlimited level of creativity in terms of how humans could be killed, but 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. 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 think through that and defend against it. There's a ceiling to the creativity of mass murder of humans here. The options are limited.

Roman Yampolskiy
They're limited by how imaginative we are. If you are that much smarter, that much more creative, you're capable of thinking across multiple domains, do 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.

Lex Fridman
So are you thinking about mass murder and destruction of human civilization or 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, majority of them are not going to be death. Most of them are going to be just things like Brave New World where the squirrels are fed dopamine and they're all doing some fun activity and the fire, the soul of humanity is lost because of the drug that's fed to it, or literally in a zoo. We're in a zoo, we're doing our thing, we're playing a game of Sims, and the actual players playing that game are AI systems. Those are all undesirable because the free will. The fire of human consciousness is dimmed through that process, but it's not killing humans. So are you thinking about that or is the biggest concern literally the extinction of humans?

Roman Yampolskiy
I think about a lot of things. So that is X-risk, existential risk, everyone's dead. There is S-risk, suffering risks, where everyone wishes they were dead. We have also idea for I-risk, ikigai risks, 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're kept alive, but we are not in control. We're not deciding anything. We're like animals in a zoo. There is, again, 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 dig into each of those X-risk, S-risk, and I-risk. So can you linger on I-risk? What is that?

Roman Yampolskiy
So Japanese concept of ikigai, you find something which allows you to make money. You are good at it and the society says we need it. So you have this awesome job. You are podcaster gives you a lot of meaning. You have a good life. I assume you're happy. That's what we want more people to find, to have. For many intellectuals, it is their occupation, which gives them a lot of meaning. I'm 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. 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. Maximize 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 king. You could be 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 the 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. We're in a simulation.

Lex Fridman
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
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
Okay, what was that paper on multi-agent value alignment?

Roman Yampolskiy
Personal universes.

Lex Fridman
So that's one of the possible outcomes, but what in general is the idea of the paper? So it's looking at multiple agents. They're human AI, like a hybrid system, whether it's humans and AIs or is it looking at humans or just intelligent agents?

Roman Yampolskiy
In order to solve 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're talking about creator of a system, owner of that AI, humanity as a whole, but we don't agree on much. There is no universally accepted ethics, morals across cultures, religions. People have individually very different preferences politically and such. So even if we somehow managed all the other aspects of it, programming those fuzzy concepts in, getting AI to follow them closely, we don't agree on what to program in. So my solution was, okay, we don't have to compromise on room temperature. 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 the multiverse theory. This is create an entire universe for you with your values.

Roman Yampolskiy
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.

Lex Fridman
So you convert the multi-agent problem into a single agent problem basically?

Roman Yampolskiy
I'm trying to do that. Yeah.

Lex Fridman
Okay. So okay, that's giving up on the value alignment problem. Well, is there any way to solve the value alignment problem where there's a bunch of humans, multiple humans, tens of humans or 8 billion humans that have 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. 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
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 a 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
Fight, 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 stuck there.

Lex Fridman
What if we want to be at tension with each other, and through that tension, we understand ourselves and we understand the world. So that's the intellectual journey we're on as a human civilization, is 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. There 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, re-engineered out. Suffering in terms of meaning, like you burn the only copy of my book, is a little harder. Even there, you can manipulate your hedonic set point, you can change defaults, you can reset. 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 as 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 positive infinity pain-pleasure axis. 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 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 maximum amount of damage, terrorism. What if someone malevolent wants on-purpose to torture all humans 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 is actually people in human history that try to literally maximize human suffering? In just studying people who have done evil in the world, it seems that they think that they're doing good and 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 be just gaining personal benefit and sacrificing others to that cause. Others we know for effect trying 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. If terrorists had access to nuclear weapons, how far would they go? Is there a limit to what they're willing to do? Your sense is there is some malevolent actors where there's no limit?

Roman Yampolskiy
There is mental diseases where people don't have empathy, don't have this human quality of understanding suffering in others.

Lex Fridman
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. There's always some sort of psychopaths, but yeah.

Lex Fridman
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, understand our molecular structure, genome. Again, a lot of times torture ends, then 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 is possible to anticipate the ways they can do it and defend against it like we do with the cybersecurity will do security systems?

Roman Yampolskiy
Right. 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 outcome long-term for humanity. The only way to win this game is not to play it.

Lex Fridman
Okay, 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? We're 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 CEO of Anthropic / DeepMind. So maybe we're two years away, which seems very soon given we don't have a working safety mechanism in place or even a

prototype for one. There are people trying to accelerate those timelines, because they feel we're not getting there quick enough.

Lex Fridman
Well, 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 you're creating 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 which 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, for me at least, if you average out over all the common human tasks, those systems are already smarter than an average human. So under that definition we have it. Shane Legg 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, 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 viruses to take over the minds of robots where the robots are executing the commands. It just seems like social engineering of humans is much more likely.

Roman Yampolskiy
That will be enough to bootstrap the whole process.

Lex Fridman
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 an inferior 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, for example?

Roman Yampolskiy
To solve pattern recognition problems of that type to have 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, that kind of stuff.

Roman Yampolskiy
We know calculators are smarter than humans in that narrow domain of addition.

Lex Fridman
Is it humans plus tools versus AGI or just human, raw human intelligence? Because humans create tools and with the tools they become more intelligent, so there's a gray area there, what it means to be human when we're measuring their intelligence.

Roman Yampolskiy
So then I think about it, I usually think human with a paper and a pencil, not human with internet and another AI helping.

Lex Fridman
Is that a fair way to think about it? Because isn't there another definition of human level intelligence that includes the tools that humans create?

Roman Yampolskiy
We create AI. So at any point you'll still just add superintelligence to human capability. That seems like cheating.

Lex Fridman
No controllable tools. There is an implied leap that you're making when AGI goes from tool to an entity that can make its own decisions. So if we define human level intelligence as everything a human can do with fully controllable tools.

Roman Yampolskiy
It seems like a hybrid of some kind. You're now doing brain computer interfaces. You're connecting it to maybe narrow AIs. Yeah, it definitely increases our capabilities.

Lex Fridman
So what's a good test to you that measures whether an artificial intelligence system has reached human level intelligence and what's a good test where it has superseded human level intelligence to reach that land of AGI?

Roman Yampolskiy
I'm old-fashioned. I like Turing tests. I have a paper where I equate passing Turing tests to solving AI complete problems because you can encode any questions about any domain into the Turing test. You don't have to talk about how was your day. You can ask anything. So the system has to be as smart as a human to pass it in a true sense.

Lex Fridman
Then you would extend that to maybe a very long conversation. I think the Alexa Prize was doing that. Basically, can you do a 20 minute, 30 minute conversation with an AI system?

Roman Yampolskiy
It has to be long enough to where you can make some meaningful decisions about capabilities, absolutely. You can brute force very short conversations.

Lex Fridman
So literally, what does that look like? Can we construct formally a test that tests for AGI?

Roman Yampolskiy
For AGI, it has to be there. I cannot give it a task I can give to a human and it cannot do it if a human can. For superintelligence, it would be superior on all such tasks, not just average performance. So go learn to drive car, go speak Chinese, play guitar. Okay, great.

Lex Fridman
I guess the follow up question, is there a test for the kind of AGI that would be susceptible to lead to S-risk or X-risk, susceptible to destroy human civilization? Is there a test for that?

Roman Yampolskiy
You can develop a test which will give you positives. If it lies to you or has those ideas, you cannot develop a test which rules them out. There is always possibility of what Bostrom calls a treacherous turn, where later on a system decides for game theoretic reasons, economic reasons to change its behavior, and we see the same with humans. It's not unique to AI. For millennia, we try developing morals, ethics, religions, lie detector tests, and then employees betray the employers, spouses betray family. It's a pretty standard thing intelligent agents sometimes do.

Lex Fridman
So is it possible to detect when a AI system is lying or deceiving you?

Roman Yampolskiy
If you know the truth and it tells you something false, you can detect that, but you cannot know in general every single time. Again, the system you're testing today may not be lying. The system you're testing today may know you are testing it, and so behaving. Later on,

after it interacts with the environment, interacts with other systems, malevolent agents learns more, it may start doing those things.

Lex Fridman
So do you think it's possible to develop a system where the creators of the system, the developers, the programmers don't know that it's deceiving them?

Roman Yampolskiy
So systems today don't have long-term planning. That is not hard. They can lie today if it helps them optimize the reward. If they realize, okay, this human will be very happy if I tell them the following, they will do it if it brings them more points. They don't have to keep track of it. It's just the right answer to this problem every single time.

Lex Fridman
At which point is somebody creating that intentionally, not unintentionally, intentionally creating an AI system that's doing long-term planning with an objective function that's defined by the AI system, not by a human?

Roman Yampolskiy
Well, some people think that if they're that smart, they're always good. They really do believe that. It just benevolence from intelligence. So they'll always want what's best for us. Some people think that they will be able to detect problem behaviors and correct them at the time when we get there. I don't think it's a good idea. I am strongly against it, but yeah, there are quite a few people who in general are so optimistic about this technology, it could do no wrong. They want it developed as soon as possible, as capable as possible.

Lex Fridman
So there's going to be people who believe the more intelligent it is, the more benevolent, and so therefore it should be the one that defines the objective function that it's optimizing when it's doing long-term planning?

Roman Yampolskiy
There are even people who say, "Okay, what's so special about humans?" Remove the gender bias, removing race bias, why is this pro-human bias? We are polluting the planet. We are, as you said, fight a lot of wars, violent. Maybe it's better if it's super intelligent, perfect society comes and replaces us. It's normal stage in the evolution of our species.

Lex Fridman
So somebody says, "Let's develop an AI system that removes the violent humans from the world." Then it turns out that all humans have violence in them or the capacity for violence and therefore all humans are removed. Yeah. Let me ask about Yann LeCun. He's somebody who you've had a few exchanges with and he's somebody who actively pushes back against

this view that AI is going to lead to destruction of human civilization, also known as AI doomerism. So in one example that he tweeted, he said, "I do acknowledge risks, but," two points, "One, open research and open source are the best ways to understand and mitigate the risks. Two, AI is not something that just happens. We build it. We have agency in what it becomes. Hence, we control the risks. We meaning humans. It's not some sort of natural phenomena that we have no control over." Can you make the case that he's right and can you try to make the case that he's wrong?

Roman Yampolskiy
I cannot make a case that he's right. He is wrong in so many ways it's difficult for me to remember all of them. He's a Facebook buddy, so I have a lot of fun having those little debates with him. So I'm trying to remember their arguments. So one, he says, we are not gifted this intelligence from aliens. We are designing it. We are making decisions about it. That's not true. It was true when we had expert systems, symbolic AI decision trees. Today, you set up parameters for a model and you water this plant. You give it data, you give it compute, and it grows. After it's finished growing into this alien plant, you start testing it to find out what capabilities it has. It takes years to figure out, even for existing models. If it's trained for six months, it'll take you two, three years to figure out basic capabilities of that system. We still discover new capabilities in systems which are already out there. So that's not the case.

Lex Fridman
So just to linger on that, so to you, the difference there is that there is some level of emergent intelligence that happens in our current approaches. So stuff that we don't hard code in.

Roman Yampolskiy
Absolutely. That's what makes it so successful. When we had to painstakingly hard code in everything, we didn't have much progress. Now, just spend more money on more compute and it's a lot more capable.

Lex Fridman
Then the question is when there is emergent intelligent phenomena, what is the ceiling of that? For you, there's no ceiling. For Yann LeCun, I think there's a ceiling that happens that we have full control over. Even if we don't understand the internals of the emergence, how the emergence happens, there's a sense that we have control and an understanding of the approximate ceiling of capability, the limits of the capability.

Roman Yampolskiy
Let's say there is a ceiling. It's not guaranteed to be at the level which is competitive with us. It may be greatly superior to ours.

Lex Fridman
So what about his statement about open research and open source are the best ways to understand and mitigate the risks?

Roman Yampolskiy
Historically, he's completely right. Open source software is wonderful. It's tested by the community, it's debugged, but we're switching from tools to agents. Now you're giving open source weapons to psychopaths. Do we want to open source nuclear weapons, biological weapons? It's not safe to give technology so powerful to those who may misalign it, even if you are successful at somehow getting it to work in the first place in a friendly manner.

Lex Fridman
The difference with nuclear weapons, current AI systems are not akin to nuclear weapons. So the idea there is you're open sourcing it at this stage that you can understand it better. Large number of people can explore the... Can understand it better. A large number of people can explore the limitation, the capabilities, explore the possible ways to keep it safe, to keep it secure, all that kind of stuff, while it's not at the stage of nuclear weapons. So nuclear weapons, there's no nuclear weapon and then there's a nuclear weapon. With AI systems, there's a gradual improvement of capability and you get to perform that improvement incrementally, and so open source allows you to study how things go wrong. I study the very process of emergence, study AI safety and those systems when there's not high level of danger, all that kind of stuff.

Roman Yampolskiy
It also sets a very wrong precedent. So we open sourced model one, model two, model three. Nothing ever bad happened, so obviously we're going to do it with model four. It's just gradual improvement.

Lex Fridman
I don't think it always works with the precedent. You're not stuck doing it the way you always did. It sets a precedent of open research and open development such that we get to learn together and then the first time there's a sign of danger, some dramatic thing happened, not a thing that destroys human civilization, but some dramatic demonstration of capability that can legitimately lead to a lot of damage, then everybody wakes up and says, "Okay, we need to regulate this. We need to come up with safety mechanism that stops this." But at this time, maybe you can educate me, but I haven't seen any illustration of significant damage done by intelligent AI systems.

Roman Yampolskiy
So I have a paper which collects accidents through history of AI and they always are proportionate to capabilities of that system. So if you have Tic-Tac-Toe playing AI, it will fail to properly play and loses the game, which it should draw trivial. Your spell checker will

misspell word, so on. I stopped collecting those because there are just too many examples of AI's failing at what they are capable of. We haven't had terrible accidents in a sense of billion people got killed. Absolutely true. But in another paper I argue that those accidents do not actually prevent people from continuing with research and actually they kind of serve like vaccines. A vaccine makes your body a little bit sick so you can handle the big disease later, much better. It's the same here. People will point out, "You know that AI accident we had where 12 people died," everyone's still here, 12 people is less than smoking kills. It's not a big deal. So we continue. So in a way it will actually be confirming that it's not that bad.

Lex Fridman
It matters how the deaths happen, whether it's literally murdered by the AI system, then one is a problem, but if it's accidents because of increased reliance on automation for example, so when airplanes are flying in an automated way, maybe the number of plane crashes increased by 17% or something, and then you're like, "Okay, do we really want to rely on automation?" I think in a case of automation airplanes, it decreased significantly. Okay, same thing with autonomous vehicles. Okay, what are the pros and cons? What are the trade-offs here? And you can have that discussion in an honest way, but I think the kind of things we're talking about here is mass scale pain and suffering caused by AI systems, and I think we need to see illustrations of that in a very small scale to start to understand that this is really damaging. Versus Clippy. Versus a tool that's really useful to a lot of people to do learning to do summarization of text, to do question-answer, all that kind of stuff to generate videos. A tool. Fundamentally a tool versus an agent that can do a huge amount of damage.

Roman Yampolskiy
So you bring up example of cars.

Lex Fridman
Yes.

Roman Yampolskiy
Cars were slowly developed and integrated. If we had no cars and somebody came around and said, "I invented this thing, it's called cars. It's awesome. It kills 100,000 Americans every year. Let's deploy it." Would we deploy that?

Lex Fridman
There'd been fear-mongering about cars for a long time. The transition from horses to cars, there's a really nice channel that I recommend people check out, Pessimist Archive that documents all the fear-mongering about technology that's happened throughout history. There's definitely been a lot of fear-mongering about cars. There's a transition period there about cars, about how deadly they are. We can try. It took a very long time for cars to

proliferate to the degree they have now. And then you could ask serious questions in terms of the miles traveled, the benefit to the economy, the benefit to the quality of life that cars do, versus the number of deaths; 30, 40,000 in the United States. Are we willing to pay that price? I think most people when they're rationally thinking, policymakers will say, "Yes." We want to decrease it from 40,000 to zero and do everything we can to decrease it. There's all kinds of policies, incentives you can create to decrease the risks with the deployment of technology. But then you have to weigh the benefits and the risks of the technology and the same thing would be done with AI.

Roman Yampolskiy
You need data, you need to know. But if I'm right and it's unpredictable, unexplainable, uncontrollable, you cannot make this decision. We're gaining $10 trillion of wealth, but we're we don't know how many people. You basically have to perform an experiment on 8 billion humans without their consent. And even if they want to give you consent, they can't because they cannot give informed consent. They don't understand those things.

Lex Fridman
Right. That happens when you go from the predictable to the unpredictable very quickly. But it's not obvious to me that AI systems would gain capabilities so quickly that you won't be able to collect enough data to study the benefits and risks.

Roman Yampolskiy
We're literally doing it. The previous model we learned about after we finished training it, what it was capable of. Let's say we stopped GPT-4 training run around human capability, hypothetically. We start training GPT- 5 and I have no knowledge of insider training runs or anything and started that point of about human and we train it for the next nine months. Maybe two months in, it becomes super intelligent. We continue training it. At the time when we start testing it, it is already a dangerous system. How dangerous? I have no idea, but never people training it.

Lex Fridman
At the training stage, but then there's a testing stage inside the company, they can start getting intuition about what the system is capable to do. You're saying that somehow from leap from GPT-4 to GPT-5 can happen, the kind of leap where GPT-4 was controllable and GPT-5 is no longer controllable and we get no insights from using GPT-4 about the fact that GPT-5 will be uncontrollable. That's the situation you're concerned about. Where there leap from N, to N plus one will be such that an uncontrollable system is created without any ability for us to anticipate that.

Roman Yampolskiy
If we had capability of ahead of the run, before the training run to register exactly what capabilities that next model will have at the end of the training run, and we accurately

guessed all of them, I would say you're right, "We can definitely go ahead with this run." We don't have the capability.

Lex Fridman
From GPT-4, you can build up intuitions about what GPT-5 will be capable of. It's just incremental progress. Even if that's a big leap in capability, it just doesn't seem like you can take a leap from a system that's helping you write emails to a system that's going to destroy human civilization. It seems like it's always going to be sufficiently incremental such that we can anticipate the possible dangers, and we're not even talking about existential risk, but just the kind of damage you can do to civilization. It seems like we'll be able to anticipate the kinds, not the exact, but the kinds of risks it might lead to and then rapidly develop defenses ahead of time and as the risks emerge.

Roman Yampolskiy
We're not talking just about capabilities specific tasks, we're talking about general capability to learn. Maybe like a child. At the time of testing and deployment, it is still not extremely capable, but as it is exposed to more data real world, it can be trained to become much more dangerous and capable.

Lex Fridman
So let's focus then on the control problem. At which point does the system become uncontrollable? Why is it the more likely trajectory for you that the system becomes uncontrollable?

Roman Yampolskiy
So, I think at some point it becomes capable of getting out of control. For game theoretic reasons, it may decide not to do anything right away and for a long time, just collect more resources, accumulate strategic advantage. Right away, it may be still young, weak super intelligence, give it a decade. It's in charge of a lot more resources, it had time to make backups. So it's not obvious to me that it will strike as soon as it can.

Lex Fridman
But can we just try to imagine this future where there's an AI system that's capable of escaping the control of humans, and then doesn't and waits? What's that look like? So one, we have to rely on that system for a lot of the infrastructure. So we'll have to give it access not just to the internet, but to the task of managing power, government, economy, this kind of stuff. And that just feels like a gradual process given the bureaucracies of all those systems involved.

Roman Yampolskiy
We've been doing it for years. Software controls all those systems, nuclear power plants, airline industry, it's all software based. Every time there is electrical outage, I can't fly anywhere for days.

Lex Fridman
But there's a difference between software and AI. So there's different kinds of software. So to give a single AI system access to the control of airlines and the control of the economy, that's not a trivial transition for humanity.

Roman Yampolskiy
No. But if it shows it is safer, in fact when it's in control, we get better results, people will demand that it was put in place.

Lex Fridman
Absolutely.

Roman Yampolskiy
And if not, it can hack the system. It can use social engineering to get access to it. That's why I said it might take some time for it to accumulate those resources.

Lex Fridman
It just feels like that would take a long time for either humans to trust it or for the social engineering to come into play. It's not a thing that happens overnight. It feels like something that happens across one or two decades.

Roman Yampolskiy
I really hope you're right, but it's not what I'm seeing. People are very quick to jump on a latest trend. Early adopters will be there before it's even deployed, buying prototypes.

Lex Fridman
Maybe the social engineering. For social engineering, AI systems don't need any hardware access. It's all software. So they can start manipulating you through social media, so on. You have AI assistants, they're going to help you manage a lot of your day to day and then they start doing social engineering. But for a system that's so capable that can escape the control of humans that created it, such a system being deployed at a mass scale and trusted by people to be deployed, it feels like that would take a lot of convincing.

Roman Yampolskiy
So, we've been deploying systems which had hidden capabilities.

Lex Fridman
Can you give an example?

Roman Yampolskiy
GPT-4. I don't know what else it's capable of, but there are still things we haven't discovered, can do. They may be trivial, proportionate with capability. I don't know it writes Chinese poetry, hypothetical, I know it does, but we haven't tested for all possible capabilities and we are not explicitly designing them. We can only rule out bugs we find. We cannot rule out bugs and capabilities because we haven't found them.

Lex Fridman
Is it possible for a system to have hidden capabilities that are orders of magnitude greater than its non- hidden capabilities? This is the thing I'm really struggling with. Where, on the surface, the thing we understand it can do doesn't seem that harmful. So even if it has bugs, even if it has hidden capabilities like Chinese poetry or generating effective viruses, software viruses, the damage that can do seems like on the same order of magnitude as the capabilities that we know about. So this idea that the hidden capabilities will include being uncontrollable is something I'm struggling with because GPT-4 on the surface seems to be very controllable.

Roman Yampolskiy
Again, we can only ask and test for things we know about. There are unknown unknowns, we cannot do it. Thinking of humans, statistics savants, right? If you talk to a person like that, you may not even realize they can multiply 20 digit numbers in their head. You have to know to ask.

Lex Fridman
So as I mentioned, just to linger on the fear of the unknown, so the Pessimist Archive has just documented, let's look at data of the past at history, there's been a lot of fear-mongering about technology. Pessimist Archive does a really good job of documenting how crazily afraid we are of every piece of technology. We've been afraid, there's a blog post where Louis Anslow who created Pessimist Archive writes about the fact that we've been fear-mongering about robots and automation for over 100 years. So why is AGI different than the kinds of technologies we've been afraid of in the past?

Roman Yampolskiy
So two things; one with wishing from tools to agents. Tools don't have negative or positive impact. People using tools do. So guns don't kill, people with guns do. Agents can make their own decisions. They can be positive or negative. A pit bull can decide to harm you. It's an agent. The fears are the same. The only difference is now we have this technology. Then they were afraid of human with robots 100 years ago, they had none. Today, every major

company in the world is investing billions to create them. Not every, but you understand what I'm saying?

Lex Fridman
Yes.

Roman Yampolskiy
It's very different.

Lex Fridman
Well, agents, it depends on what you mean by the word, "Agents." All those companies are not investing in a system that has the kind of agency that's implied by in the fears, where it can really make decisions on their own, that have no human in the loop.

Roman Yampolskiy
They are saying they're building super intelligence and have a Super Alignment Team. You don't think they're trying to create a system smart enough to be an independent agent? Under that definition?

Lex Fridman
I have not seen evidence of it. I think a lot of it is a marketing kind of discussion about the future and it's a mission about the kind of systems we can create in the long term future. But in the short term, the kind of systems they're creating falls fully within the definition of narrow AI. These are tools that have increasing capabilities, but they just don't have a sense of agency, or consciousness, or self-awareness or ability to deceive at scales that would be required to do mass scale suffering and murder of humans.

Roman Yampolskiy
Those systems are well beyond narrow AI. If you had to list all the capabilities of GPT-4, you would spend a lot of time writing that list.

Lex Fridman
But agency is not one of them.

Roman Yampolskiy
Not yet. But do you think any of those companies are holding back because they think it may be not safe? Or are they developing the most capable system they can given the resources and hoping they can control and monetize?

Lex Fridman
Control and monetize. Hoping they can control and monetize. So you're saying if they could press a button, and create an agent that they no longer control, that they have to ask nicely,

a thing that lives on a server, across huge number of computers, you're saying that they would push for the creation of that kind of system?

Roman Yampolskiy
I mean, I can't speak for other people, for all of them. I think some of them are very ambitious. They're fundraising trillions, they talk about controlling the light corner of the universe. I would guess that they might.

Lex Fridman
Well, that's a human question, whether humans are capable of that. Probably, some humans are capable of that. My more direct question, if it's possible to create such a system, have a system that has that level of agency. I don't think that's an easy technical challenge. It doesn't feel like we're close to that. A system that has the kind of agency where it can make its own decisions and deceive everybody about them. The current architecture we have in machine learning and how we train the systems, how to deploy the systems and all that, it just doesn't seem to support that kind of agency.

Roman Yampolskiy
I really hope you are right. I think the scaling hypothesis is correct. We haven't seen diminishing returns. It used to be we asked how long before AGI, now we should ask how much until AGI, it's $1 trillion today it's $1 billion next year, it's $1 million in a few years.

Lex Fridman
Don't you think it's possible to basically run out of trillions? So is this constrained by compute?

Roman Yampolskiy
Compute gets cheaper every day, exponentially.

Lex Fridman
But then it becomes a question of decades versus years.

Roman Yampolskiy
If the only disagreement is that it will take decades, not years for everything I'm saying to materialize, then I can go with that.

Lex Fridman
But if it takes decades, then the development of tools for AI safety then becomes more and more realistic. So I guess the question is, I have a fundamental belief that humans when faced with danger, can come up with ways to defend against that danger. And one of the big problems facing AI safety currently, for me, is that there's not clear illustrations of what that danger looks like. There's no illustrations of AI systems doing a lot of damage, and so it's

unclear what you're defending against. Because currently it's a philosophical notions that, yes, it's possible to imagine AI systems that take control of everything and then destroy all humans. It's also a more formal mathematical notion that you talk about that it's impossible to have a perfectly secure system. You can't prove that a program of sufficient complexity is completely safe, and perfect and know everything about it, yes, but when you actually just pragmatically look how much damage have the AI systems done and what kind of damage, there's not been illustrations of that. Even in the autonomous weapon systems, there's not been mass deployments of autonomous weapon systems, luckily. The automation in war currently is very limited, that the automation is at the scale of individuals versus at the scale of strategy and planning. I think one of the challenges here is where is the dangers and the intuition the [inaudible 00:54:40] and others have is, let's keep in the open building AI systems until the dangers start rearing their heads and they become more explicit, they start being case studies, illustrative case studies that show exactly how the damage by AD systems is done, then regulation can step in. Then brilliant engineers can step up, and we can have Manhattan style projects that defend against such systems. That's kind of the notion. And I guess, a tension with that is the idea that for you, we need to be thinking about that now, so that we're ready, because we'll have not much time once the systems are deployed. Is that true?

Roman Yampolskiy
So, there is a lot to unpack here. There is a partnership on AI, a conglomerate of many large corporations. They have a database of AI accidents they collect. I contributed a lot to that database. If we so far made almost no progress in actually solving this problem, not patching it, not again, lipstick on a pig kind of solutions, why would we think we'll do better when we're closer to the problem?

Lex Fridman
All the things you mentioned are serious concerns measuring the amount of harm. So benefit versus risk there is difficult. But to you, the sense is already the risk has superseded the benefit?

Roman Yampolskiy
Again, I want to be perfectly clear, I love AI, I love technology. I'm a computer scientist. I have PhD in engineering. I work at an engineering school. There is a huge difference between we need to develop mar AI systems, super intelligent in solving specific human problems like protein folding and let's create super intelligent machine guards that will decide what to do with us. Those are not the same. I am against the super intelligence in general sense with no undue burden.

Lex Fridman
So do you think the teams that are able to do the AI safety on the kind of narrow AI risks that you've mentioned, are those approaches going to be at all productive towards leading to approaches of doing AI safety on AGI? Or is it just a fundamentally different part?

Roman Yampolskiy
Partially, but we don't scale for narrow AI for deterministic systems. You can test them, you have edge cases. You know what the answer should look like, the right answers. For general systems, you have infinite test surface, you have no edge cases. You cannot even know what to test for. Again, the unknown unknowns are underappreciated by people looking at this problem. You are always asking me, "How will it kill everyone? How will it will fail?" The whole point is if I knew it, I would be super intelligent and despite what you might think, I'm not.

Lex Fridman
So to you, the concern is that we would not be able to see early signs of an uncontrollable system.

Roman Yampolskiy
It is a master at deception. Sam tweeted about how great it is at persuasion and we see it ourselves, especially now with voices with maybe kind of flirty, sarcastic female voices. It's going to be very good at getting people to do things.

Lex Fridman
But see, I'm very concerned about system being used to control the masses. But in that case, the developers know about the kind of control that's happening. You're more concerned about the next stage where even the developers don't know about the deception.

Roman Yampolskiy
Correct. I don't think developers know everything about what they are creating. They have lots of great knowledge, we're making progress on explaining parts of a network. We can understand, "Okay, this note get excited, then this input is presented, this cluster of notes." But we're nowhere near close to understanding the full picture, and I think it's impossible. You need to be able to survey an explanation. The size of those models prevents a single human from absorbing all this information, even if provided by the system. So either we're getting model as an explanation for what's happening and that's not comprehensible to us or we're getting compressed explanation, [inaudible 00:59:01] compression, where here, "Top 10 reasons you got fired." It's something, but it's not a full picture.

Lex Fridman
You've given elsewhere an example of a child and everybody, all humans try to deceive, they try to lie early on in their life. I think we'll just get a lot of examples of deceptions from large language models or AI systems. They're going to be kind of shady, or they'll be pretty good, but we'll catch them off guard. We'll start to see the kind of momentum towards developing increasing deception capabilities and that's when you're like, "Okay, we need to do some kind of alignment that prevents deception." But, if you support open source, then you can have open source models that have some level of deception you can start to explore on a large scale, how do we stop it from being deceptive? Then there's a more explicit, pragmatic kind of problem to solve. How do we stop AI systems from trying to optimize for deception? That's an example.

Roman Yampolskiy
So there is a paper, I think it came out last week by Dr Park et al, from MIT I think, and they showed that models already showed successful deception in what they do. My concern is not that they lie now, and we need to catch them and tell them, "Don't lie." My concern is that once they are capable and deployed, they will later change their mind. Because what unrestricted learning allows you to do. Lots of people grow up maybe in the religious family, they read some new books and they turn in their religion. That's a treacherous turn in humans. If you learn something new about your colleagues, maybe you'll change how you react to that.

Lex Fridman
Yeah, the treacherous turn. If we just mention humans, Stalin and Hitler, there's a turn. Stalin's a good example. He just seems like a normal communist follower of Lenin until there's a turn. There's a turn of what that means in terms of when he has complete control, what the execution of that policy means and how many people get to suffer.

Roman Yampolskiy
And you can't say they are not rational. The rational decision changes based on your position. When you are under the boss, the rational policy may be to be following orders and being honest. When you become a boss, rational policy may shift.

Lex Fridman
Yeah, and by the way, a lot of my disagreements here is just playing Devil's Advocate to challenge your ideas and to explore them together. So one of the big problems here in this whole conversation is human civilization hangs in the balance and yet everything's unpredictable. We don't know how these systems will look like-

Roman Yampolskiy
The robots are coming.

Lex Fridman
There's a refrigerator making a buzzing noise.

Roman Yampolskiy
Very menacing. Very menacing. So every time I'm about to talk about this topic, things start to happen. My flight yesterday was canceled without possibility to re-book. I was giving a talk at Google in Israel and three cars, which were supposed to take me to the talk could not. I'm just saying.

Lex Fridman
I mean

Roman Yampolskiy
I like AI's. I, for one welcome our overlords.

Lex Fridman
There's a degree to which we... I mean it is very obvious as we already have, we've increasingly given our life over to software systems. And then it seems obvious given the capabilities of AI that are coming, that we'll give our lives over increasingly to AI systems. Cars will drive themselves, refrigerator eventually will optimize what I get to eat. And, as more and more out of our lives are controlled or managed by AI assistants, it is very possible that there's a drift. I mean, I personally am concerned about non-existential stuff, the more near term things. Because before we even get to existential, I feel like there could be just so many Brave New World type of situations. You mentioned the term, "Behavioral drift." It's the slow boiling that I'm really concerned about as we give our lives over to automation, that our minds can become controlled by governments, by companies, or just in a distributed way. There's a drift. Some aspect of our human nature gives ourselves over to the control of AI systems and they, in an unintended way just control how we think. Maybe there'll be a herd-like mentality in how we think, which will kill all creativity and exploration of ideas, the diversity of ideas, or much worse. So it's true, it's true. But a lot of the conversation I'm having with you now is also kind of wondering almost at a technical level, how can AI escape control? What would that system look like? Because it, to me, is terrifying and fascinating. And also fascinating to me is maybe the optimistic notion it's possible to engineer systems that defend against that. One of the things you write a lot about in your book is verifiers. So, not humans. Humans are also verifiers. But software systems that look at AI systems, and help you understand, "This thing is getting real weird." Help you analyze those systems. So maybe this is a good time to talk about verification. What is this beautiful notion of verification?

Roman Yampolskiy
My claim is, again, that there are very strong limits in what we can and cannot verify. A lot of times when you post something on social media, people go, "Oh, I need citation to a peer

reviewed article." But what is a peer reviewed article? You found two people in a world of hundreds of thousands of scientists who said, "Ah, whatever, publish it. I don't care." That's the verifier of that process. When people say, "Oh, it's formally verified software or mathematical proof," we accept something close to 100% chance of it being free of all problems. But you actually look at research, software is full of bugs, old mathematical theorems, which have been proven for hundreds of years have been discovered to contain bugs, on top of which we generate new proofs and now we have to redo all that. So, verifiers are not perfect. Usually, they are either a single human or communities of humans and it's basically kind of like a democratic vote. Community of mathematicians agrees that this proof is correct, mostly correct. Even today, we're starting to see some mathematical proofs as so complex, so large that mathematical community is unable to make a decision. It looks interesting, it looks promising, but they don't know. They will need years for top scholars to study to figure it out. So of course, we can use AI to help us with this process, but AI is a piece of software which needs to be verified.

Lex Fridman
Just to clarify, so verification is the process of something is correct, it is the formal, and mathematical proof, where's a statement, and a series of logical statements that prove that statement to be correct, which is a theorem. And you're saying it gets so complex that it's possible for the human verifiers, the human beings that verify that the logical step, there's no bugs in it becomes impossible. So, it's nice to talk about verification in this most formal, most clear, most rigorous formulation of it, which is mathematical proofs.

Roman Yampolskiy
Right. And for AI we would like to have that level of confidence for very important mission-critical software controlling satellites, nuclear power plants. For small, deterministic programs We can do this, we can check that code verifies its mapping to the design. Whatever software engineers intended, was correctly implemented. But we don't know how to do this for software which keeps learning, self-modifying, rewriting its own code. We don't know how to prove things about the physical world, states of humans in the physical world. So there are papers coming out now and I have this beautiful one, "Towards Guaranteed Safe AI." Very cool papers, some of the best [inaudible 01:07:54] I ever seen. I think there is multiple Turing Award winners that is quite... You can have this one and one just came out kind of similar, "Managing Extreme-" ... one just came out kind of similar, managing extremely high risks. So, all of them expect this level of proof, but I would say that we can get more confidence with more resources we put into it. But at the end of the day, we're still as reliable as the verifiers. And you have this infinite regress of verifiers. The software used to verify a program is itself a piece of program. If aliens give us well-aligned super intelligence, we can use that to create our own safe AI. But it's a catch-22. You need to have already proven to be safe system to verify this new system of equal or greater complexity.

Lex Fridman
You just mentioned this paper, Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems. Like you mentioned, it's like a who's who. Josh Tenenbaum, Yoshua Bengio, Stuart Russell, Max Tegmark, and many other brilliant people. The page you have it open on, "There are many possible strategies for creating safety specifications. These strategies can roughly be placed on a spectrum, depending on how much safety it would grant if successfully implemented. One way to do this is as follows," and there's a set of levels. From Level 0, "No safety specification is used," to Level 7, "The safety specification completely encodes all things that humans might want in all contexts." Where does this paper fall short to you?

Roman Yampolskiy
So, when I wrote a paper, Artificial Intelligence Safety Engineering, which kind of coins the term AI safety, that was 2011. We had 2012 conference, 2013 journal paper. One of the things I proposed, let's just do formal verifications on it. Let's do mathematical formal proofs. In the follow-up work, I basically realized it will still not get us a hundred percent. We can get 99.9, we can put more resources exponentially and get closer, but we never get to a hundred percent. If a system makes a billion decisions a second, and you use it for a hundred years, you're still going to deal with a problem. This is wonderful research. I'm so happy they're doing it. This is great, but it is not going to be a permanent solution to that problem.

Lex Fridman
Just to clarify, the task of creating an AI verifier is what? Is creating a verifier that the AI system does exactly as it says it does, or it sticks within the guardrails that it says it must?

Roman Yampolskiy
There are many, many levels. So, first you're verifying the hardware in which it is run. You need to verify communication channel with the human. Every aspect of that whole world model needs to be verified. Somehow, it needs to map the world into the world model, map and territory differences. How do I know internal states of humans? Are you happy or sad? I can't tell. So, how do I make proofs about real physical world? Yeah, I can verify that deterministic algorithm follows certain properties, that can be done. Some people argue that maybe just maybe two plus two is not four. I'm not that extreme. But once you have sufficiently large proof over sufficiently complex environment, the probability that it has zero bugs in it is greatly reduced. If you keep deploying this a lot, eventually you're going to have a bug anyways.

Lex Fridman
There's always a bug.

Roman Yampolskiy
There is always a bug. And the fundamental difference is what I mentioned. We're not dealing with cybersecurity. We're not going to get a new credit card, new humanity.

Lex Fridman
So, this paper is really interesting. You said 2011, Artificial Intelligence, Safety Engineering. Why Machine Ethics is a Wrong Approach. The grand challenge you write of AI safety engineering, "We propose the problem of developing safety mechanisms for self-improving systems." Self-improving systems. By the way, that's an interesting term for the thing that we're talking about. Is self-improving more general than learning? Self-improving, that's an interesting term.

Roman Yampolskiy
You can improve the rate at which you are learning, you can become more efficient, meta-optimizer.

Lex Fridman
The word self, it's like self replicating, self improving. You can imagine a system building its own world on a scale and in a way that is way different than the current systems do. It feels like the current systems are not self-improving or self-replicating or self-growing or self-spreading, all that kind of stuff. And once you take that leap, that's when a lot of the challenges seems to happen because the kind of bugs you can find now seems more akin to the current normal software debugging kind of process. But whenever you can do self-replication and arbitrary self-improvement, that's when a bug can become a real problem, real fast. So, what is the difference to you between verification of a non-self-improving system versus a verification of a self-improving system?

Roman Yampolskiy
So, if you have fixed code for example, you can verify that code, static verification at the time, but if it will continue modifying it, you have a much harder time guaranteeing that important properties of that system have not been modified than the code changed.

Lex Fridman
Is it even doable?

Roman Yampolskiy
No.

Lex Fridman
Does the whole process of verification just completely fall apart?

Roman Yampolskiy
It can always cheat. It can store parts of its code outside in the environment. It can have extended mind situations. So, this is exactly the type of problems I'm trying to bring up.

Lex Fridman
What are the classes of verifiers that you read about in the book? Is there interesting ones that stand out to you? Do you have some favorites?

Roman Yampolskiy
I like Oracle types where you just know that it's right. Turing likes Oracle machines. They know the right answer. How? Who knows? But they pull it out from somewhere, so you have to trust them. And that's a concern I have about humans in a world with very smart machines. We experiment with them. We see after a while, okay, they've always been right before, and we start trusting them without any verification of what they're saying.

Lex Fridman
Oh, I see. That we kind of build Oracle verifiers or rather we build verifiers we believe to be Oracles and then we start to, without any proof, use them as if they're Oracle verifiers.

Roman Yampolskiy
We remove ourselves from that process. We're not scientists who understand the world. We are humans who get new data presented to us.

Lex Fridman
Okay, one really cool class of verifiers is a self verifier. Is it possible that you somehow engineer into AI system, the thing that constantly verifies itself

Roman Yampolskiy
Preserved portion of it can be done, but in terms of mathematical verification, it's kind of useless. You saying you are the greatest guy in the world because you are saying it, it's circular and not very helpful, but it's consistent. We know that within that world, you have verified that system. In a paper, I try to brute force all possible verifiers. It doesn't mean that this one particularly important to us.

Lex Fridman
But what about self-doubt? The kind of verification where you said, you say, or I say I'm the greatest guy in the world. What about a thing which I actually have is a voice that is constantly extremely critical. So, engineer into the system a constant uncertainty about self, a constant doubt.

Roman Yampolskiy
Any smart system would have doubt about everything. You not sure if what information you are given is true. If you are subject to manipulation, you have this safety and security mindset.

Lex Fridman
But I mean, you have doubt about yourself. The AI systems that has a doubt about whether the thing is doing is causing harm is the right thing to be doing. So, just a constant doubt about what it's doing because it's hard to be a dictator full of doubt.

Roman Yampolskiy
I may be wrong, but I think Stuart Russell's ideas are all about machines which are uncertain about what humans want and trying to learn better and better what we want. The problem of course is we don't know what we want and we don't agree on it.

Lex Fridman
Yeah, but uncertainty. His idea is that having that self-doubt uncertainty in AI systems, engineering into AI systems, is one way to solve the control problem.

Roman Yampolskiy
It could also backfire. Maybe you're uncertain about completing your mission. Like I am paranoid about your cameras not recording right now. So, I would feel much better if you had a secondary camera, but I also would feel even better if you had a third and eventually I would turn this whole world into cameras pointing at us, making sure we're capturing this.

Lex Fridman
No, but wouldn't you have a meta concern like that you just stated, that eventually there'd be way too many cameras? So, you would be able to keep zooming on the big picture of your concerns.

Roman Yampolskiy
So, it's a multi-objective optimization. It depends, how much I value capturing this versus not destroying the universe.

Lex Fridman
Right, exactly. And then you will also ask about, "What does it mean to destroy the universe? And how many universes are?" And you keep asking that question, but that doubting yourself would prevent you from destroying the universe because you're constantly full of doubt. It might affect your productivity.

Roman Yampolskiy
You might be scared to do anything.

Lex Fridman
Just scared to do anything.

Roman Yampolskiy
Mess things up.

Lex Fridman
Well, that's better. I mean, I guess the question, is it possible to engineer that in? I guess your answer would be yes, but we don't know how to do that and we need to invest a lot of effort into figuring out how to do that, but it's unlikely. Underpinning a lot of your writing is this sense that we're screwed, but it just feels like it's an engineering problem. I don't understand why we're screwed. Time and time again, humanity has gotten itself into trouble and figured out a way to get out of the trouble.

Roman Yampolskiy
We are in a situation where people making more capable systems just need more resources. They don't need to invent anything, in my opinion. Some will disagree, but so far at least I don't see diminishing returns. If you have 10X compute, you will get better performance. The same doesn't apply to safety. If you give MIRI or any other organization 10 times the money, they don't output 10 times the safety. And the gap between capabilities and safety becomes bigger and bigger all the time. So, it's hard to be completely optimistic about our results here. I can name 10 excellent breakthrough papers in machine learning. I would struggle to name equally important breakthroughs in safety. A lot of times a safety paper will propose a toy solution and point out 10 new problems discovered as a result. It's like this fractal. You're zooming in and you see more problems and it's infinite in all directions.

Lex Fridman
Does this apply to other technologies or is this unique to AI, where safety is always lagging behind?

Roman Yampolskiy
I guess we can look at related technologies with cybersecurity, right? We did manage to have banks and casinos and Bitcoin, so you can have secure narrow systems which are doing okay. Narrow attacks on them fail, but you can always go outside of a box. So, if I can hack your Bitcoin, I can hack you. So there is always something, if I really want it, I will find a different way. We talk about guardrails for AI. Well, that's a fence. I can dig a tunnel under it, I can jump over it, I can climb it, I can walk around it. You may have a very nice guardrail, but in a real world it's not a permanent guarantee of safety. And again, this is a fundamental difference. We are not saying we need to be 90% safe to get those trillions of dollars of benefit. We need to be a hundred percent indefinitely or we might lose the principle.

Lex Fridman
So, if you look at just humanity as a set of machines, is the machinery of AI safety conflicting with the machinery of capitalism.

Roman Yampolskiy
I think we can generalize it to just prisoners' dilemma in general. Personal self-interest versus group interest. The incentives are such that everyone wants what's best for them. Capitalism obviously has that tendency to maximize your personal gain, which does create this race to the bottom. I don't have to be a lot better than you, but if I'm 1% better than you, I'll capture more of the profits, so it's worth for me personally to take the risk even if society as a whole will suffer as a result.

Lex Fridman
But capitalism has created a lot of good in this world. It's not clear to me that AI safety is not aligned with the function of capitalism, unless AI safety is so difficult that it requires the complete halt of the development, which is also a possibility. It just feels like building safe systems should be the desirable thing to do for tech companies.

Roman Yampolskiy
Right. Look at governance structures. When you have someone with complete power, they're extremely dangerous. So, the solution we came up with is break it up. You have judicial, legislative, executive. Same here, have narrow AI systems, work on important problems. Solve immortality. It's a biological problem we can solve similar to how progress was made with protein folding, using a system which doesn't also play chess. There is no reason to create super intelligent system to get most of the benefits we want from much safer narrow systems.

Lex Fridman
It really is a question to me whether companies are interested in creating anything but narrow AI. I think when term AGI is used by tech companies, they mean narrow AI. They mean narrow AI with amazing capabilities. I do think that there's a leap between narrow AI with amazing capabilities, with superhuman capabilities and the kind of self-motivated agent-like AGI system that we're talking about. I don't know if it's obvious to me that a company would want to take the leap to creating an AGI that it would lose control of because then you can't capture the value from that system.

Roman Yampolskiy
The bragging rights, but being-

Lex Fridman
That's a different-

Roman Yampolskiy
... first, that is the same humans who are in charge of those systems.

Lex Fridman
That's a human thing. That's so that jumps from the incentives of capitalism to human nature. And so the question is whether human nature will override the interest of the company. So, you've mentioned slowing or halting progress. Is that one possible solution? Are you proponent of pausing development of AI, whether it's for six months or completely?

Roman Yampolskiy
The condition would be not time, but capabilities. Pause until you can do X, Y, Z. And if I'm right and you cannot, it's impossible, then it becomes a permanent ban. But if you're right, and it's possible, so as soon as you have those safety capabilities, go ahead.

Lex Fridman
Right. Is there any actual explicit capabilities that you can put on paper, that we as a human civilization could put on paper? Is it possible to make it explicit like that versus kind of a vague notion of just like you said, it's very vague. We want AI systems to do good and want them to be safe. Those are very vague notions. Is there more formal notions?

Roman Yampolskiy
So, when I think about this problem, I think about having a toolbox I would need. Capabilities such as explaining everything about that system's design and workings, predicting not just terminal goal, but all the intermediate steps of a system. Control in terms of either direct control, some sort of a hybrid option, ideal advisor. It doesn't matter which one you pick, but you have to be able to achieve it. In a book we talk about others, verification is another very important tool. Communication without ambiguity, human language is ambiguous. That's another source of danger. So, basically there is a paper we published in ACM surveys, which looks at about 50 different impossibility results, which may or may not be relevant to this problem, but we don't have enough human resources to investigate all of them for relevance to AI safety. The ones I mentioned to you, I definitely think would be handy, and that's what we see AI safety researchers working on. Explainability is a huge one. The problem is that it's very hard to separate capabilities work from safety work. If you make good progress in explainability, now the system itself can engage in self-improvement much easier, increasing capability greatly. So, it's not obvious that there is any research which is pure safety work without disproportionate increasing capability and danger.

Lex Fridman
Explainability is really interesting. Why is that connected to you to capability? If it's able to explain itself well, why does that naturally mean that it's more capable?

Roman Yampolskiy
Right now, it's comprised of weights and a neural network. If it can convert it to manipulatable code, like software, it's a lot easier to work in self-improvement.

Lex Fridman
I see. So, it increases-

Roman Yampolskiy
You can do intelligent design instead of evolutionary, gradual descent.

Lex Fridman
Well, you could probably do human feedback, human alignment more effectively if it's able to be explainable. If it's able to convert the weights into human understandable form, then you could probably have humans interact with it better. Do you think there's hope that we can make AI systems explainable?

Roman Yampolskiy
Not completely. So, if they are sufficiently large, you simply don't have the capacity to comprehend what all the trillions of connections represent. Again, you can obviously get a very useful explanation which talks about the top most important features which contribute to the decision, but the only true explanation is the model itself.

Lex Fridman
Deception could be part of the explanation, right? So you can never prove that there's some deception in the networks explaining itself.

Roman Yampolskiy
Absolutely. And you can probably have targeted deception where different individuals will understand explanation in different ways based on their cognitive capability. So, while what you're saying may the same and true in some situations, ours will be deceived by it.

Lex Fridman
So, it's impossible for an AI system to be truly fully explainable in the way that we mean honestly and [inaudible 01:27:57]-

Roman Yampolskiy
Again, at the extreme. The systems which are narrow and less complex could be understood pretty well.

Lex Fridman
If it's impossible to be perfectly explainable, is there a hopeful perspective on that? It's impossible to be perfectly explainable, but you can explain most of the important stuff? You can ask a system, "What are the worst ways you can hurt humans?" And it'll answer honestly.

Roman Yampolskiy
Any work in a safety direction right now seems like a good idea because we are not slowing down. I'm not for a second thinking that my message or anyone else's will be heard and will be a sane civilization, which decides not to kill itself by creating its own replacements.

Lex Fridman
The pausing of development is an impossible thing for you.

Roman Yampolskiy
Again, it's always limited by either geographic constraints, pause in US, pause in China. So, there are other jurisdictions as the scale of a project becomes smaller. So, right now it's like Manhattan Project scale in terms of costs and people. But if five years from now, compute is available on a desktop to do it, regulation will not help. You can't control it as easy. Any kid in the garage can train a model. So, a lot of it is, in my opinion, just safety theater, security theater where we saying, "Oh, it's illegal to train models so big." Okay.

Lex Fridman
So okay, that's security theater and is government regulation also security theater?

Roman Yampolskiy
Given that a lot of the terms are not well-defined and really cannot be enforced in real life. We don't have ways to monitor training runs meaningfully life while they take place. There are limits to testing for capabilities I mentioned, so a lot of it cannot be enforced. Do I strongly support all that regulation? Yes, of course. Any type of red tape will slow it down and take money away from compute towards lawyers.

Lex Fridman
Can you help me understand, what is the hopeful path here for you solution wise out of this? It sounds like you're saying AI systems in the end are unverifiable, unpredictable. As the book says, unexplainable, uncontrollable.

Roman Yampolskiy
That's the big one.

Lex Fridman
Uncontrollable, and all the other uns just make it difficult to avoid getting to the uncontrollable, I guess. But once it's uncontrollable, then it just goes wild. Surely there are

solutions. Humans are pretty smart. What are possible solutions? If you are a dictator of the world, what do we do?

Roman Yampolskiy
The smart thing is not to build something you cannot control, you cannot understand. Build what you can and benefit from it. I'm a big believer in personal self-interest. A lot of guys running those companies are young, rich people. What do they have to gain beyond billions they already have financially, right? It's not a requirement that they press that button. They can easily wait a long time. They can just choose not to do it and still have amazing life. In history, a lot of times if you did something really bad, at least you became part of history books. There is a chance in this case there won't be any history.

Lex Fridman
So, you're saying the individuals running these companies should do some soul-searching and what? And stop development?

Roman Yampolskiy
Well, either they have to prove that, of course it's possible to indefinitely control godlike, super-intelligent machines by humans and ideally let us know how, or agree that it's not possible and it's a very bad idea to do it. Including for them personally and their families and friends and capital.

Lex Fridman
What do you think the actual meetings inside these companies look like? Don't you think all the engineers... Really it is the engineers that make this happen. They're not like automatons. They're human beings. They're brilliant human beings. They're non-stop asking, how do we make sure this is safe?

Roman Yampolskiy
So again, I'm not inside. From outside, it seems like there is a certain filtering going on and restrictions and criticism and what they can say. And everyone who was working in charge of safety and whose responsibility it was to protect us said, "You know what? I'm going home." So, that's not encouraging.

Lex Fridman
What do you think the discussion inside those companies look like? You're developing, you're training GPT-V, you're training Gemini, you're training Claude and Grok. Don't you think they're constantly, like underneath it, maybe it's not made explicit, but you're constantly sort of wondering where's the system currently stand? Where are the possible unintended consequences? Where are the limits? Where are the bugs? The small and the big bugs? That's the constant thing that engineers are worried about. I think super alignment is not quite the same as the kind of thing I'm referring to with engineers are worried about. Super

alignment is saying, "For future systems that we don't quite yet have, how do we keep them safe?" You are trying to be a step ahead. It's a different kind of problem because it is almost more philosophical. It's a really tricky one because you're trying to prevent future systems from escaping control of humans. I don't think there's been... Man, is there anything akin to it in the history of humanity? I don't think so, right?

Roman Yampolskiy
Climate change.

Lex Fridman
But there's a entire system which is climate, which is incredibly complex, which we have only tiny control of, right? It's its own system. In this case, we're building the system. So, how do you keep that system from becoming destructive? That's a really different problem than the current meetings that companies are having where the engineers are saying, "Okay, how powerful is this thing? How does it go wrong? And as we train GPT-V and train up future systems, where are the ways that can go wrong?" Don't you think all those engineers are constantly worrying about this, thinking about this? Which is a little bit different than the super alignment team that's thinking a little bit farther into the future.

Roman Yampolskiy
Well, I think a lot of people who historically worked on AI never considered what happens when they succeed. Stuart Russell speaks beautifully about that. Let's look, okay, maybe superintelligence is too futuristic. We can develop practical tools for it. Let's look at software today. What is the state of safety and security of our user software? Things we give to millions of people? There is no liability. You click, "I agree." What are you agreeing to? Nobody knows. Nobody reads. But you're basically saying it will spy on you, corrupt your data, kill your firstborn, and you agree and you're not going to sue the company. That's the best they can do for mundane software, word processor, tax software. No liability, no responsibility. Just as long as you agree not to sue us, you can use it. If this is a state of the art in systems which are narrow accountants, stable manipulators, why do we think we can do so much better with much more complex systems across multiple domains in the environment with malevolent actors? With again, self-improvement with capabilities exceeding those of humans thinking about it.

Lex Fridman
I mean, the liability thing is more about lawyers than killing firstborns. But if Clippy actually killed the child, I think lawyers aside, it would end Clippy and the company that owns Clippy. So, it's not so much about... There's two points to be made. One is like, man, current software systems are full of bugs and they could do a lot of damage and we don't know what, they're unpredictable. There's so much damage they could possibly do. And then we kind of live in this blissful illusion that everything is great and perfect and it works. Nevertheless, it still somehow works.

Roman Yampolskiy
In many domains, we see car manufacturing, drug development, the burden of proof is on a manufacturer of product or service to show their product or is safe. It is not up to the user to prove that there are problems. They have to do appropriate safety studies. We have to get government approval for selling the product and they're still fully responsible for what happens. We don't see any of that here. They can deploy whatever they want and I have to explain how that system is going to kill everyone. I don't work for that company. You have to explain to me how it's definitely cannot mess up.

Lex Fridman
That's because it's the very early days of such a technology. Government regulation is lagging behind. They're really not tech-savvy. A regulation of any kind of software. If you look at Congress talking about social media and whenever Mark Zuckerberg and other CEOs show up, the cluelessness that Congress has about how technology works is incredible. It's heartbreaking, honestly

Roman Yampolskiy
I agree completely, but that's what scares me. The response is, "When they start to get dangerous, we'll really get it together. The politicians will pass the right laws, engineers will solve the right problems." We are not that good at many of those things, we take forever. And we are not early. We are two years away according to prediction markets. This is not a biased CEO fund-raising. This is what smartest people, super forecasters are thinking of this problem.

Lex Fridman
I'd like to push back about those... I wonder what those prediction markets are about, how they define AGI. That's wild to me. And I want to know what they said about autonomous vehicles because I've heard a lot of experts and financial experts talk about autonomous vehicles and how it's going to be a multi-trillion dollar industry and all this kind of stuff, and it's...

Roman Yampolskiy
A small font, but if you have good vision, maybe you can zoom in on that and see a prediction dates in the description.

Lex Fridman
Oh, there's a plot.

Roman Yampolskiy
I have a large one if you're interested.

Lex Fridman
I guess my fundamental question is how often they write about technology. I definitely do-

Roman Yampolskiy
There are studies on their accuracy rates and all that. You can look it up. But even if they're wrong, I'm just saying this is right now the best we have, this is what humanity came up with as the predicted date.

Lex Fridman
But again, what they mean by AGI is really important there. Because there's the non-agent like AGI, and then there's an agent like AGI, and I don't think it's as trivial as a wrapper. Putting a wrapper around, one has lipstick and all it takes is to remove the lipstick. I don't think it's that trivial.

Roman Yampolskiy
You may be completely right, but what probability would you assign it? You may be 10% wrong, but we're betting all of humanity on this distribution. It seems irrational.

Lex Fridman
Yeah, it's definitely not like 1 or 0%. Yeah. What are your thoughts, by the way, about current systems, where they stand? GPT-4.0, Claude 2, Grok, Gemini. On the path to super intelligence, to agent-like super intelligence, where are we?

Roman Yampolskiy
I think they're all about the same. Obviously there are nuanced differences, but in terms of capability, I don't see a huge difference between them. As I said, in my opinion, across all possible tasks, they exceed performance of an average person. I think they're starting to be better than an average masters student at my university, but they still have very big limitations. If the next model is as improved as GPT-4 versus GPT-3, we may see something very, very, very capable.

Lex Fridman
What do you feel about all this? I mean, you've been thinking about AI safety for a long, long time. And at least for me, the leaps, I mean, it probably started with... AlphaZero was mind-blowing for me, and then the breakthroughs with LLMs, even GPT-II, but just the breakthroughs on LLMs, just mind-blowing to me. What does it feel like to be living in this day and age where all this talk about AGIs feels like it actually might happen, and quite soon, meaning within our lifetime? What does it feel like?

Roman Yampolskiy
So, when I started working on this, it was pure science fiction. There was no funding, no journals, no conferences known in academia would dare to touch anything with the word

singularity in it. And I was pretty tenured at the time, so I was pretty dumb. Now you see Turing Award winners publishing in science about how far behind we are according to them in addressing this problem. So, it's definitely a change. It's difficult to keep up. I used to be able to read every paper on AI safety. Then I was able to read the best ones. Then the titles, and now I don't even know what's going on. By the time this interview is over, they probably had GPT-VI released, and I have to deal with that when I get back home. ... GPT6 released and I have to deal with that when I get back home. So it's interesting. Yes, there is now more opportunities. I get invited to speak to smart people.

Lex Fridman
By the way, I would've talked to you before any of this. This is not like some trend of AI... To me, we're still far away. So just to be clear, we're still far away from AGI, but not far away in the sense... Relative to the magnitude of impact it can have, we're not far away, and we weren't far away 20 years ago because the impact AGI can have is on a scale of centuries. It can end human civilization or it can transform it. So this discussion about one or two years versus one or two decades or even a hundred years is not as important to me, because we're headed there. This is like a human, civilization scale question. So this is not just a hot topic.

Roman Yampolskiy
It is the most important problem we'll ever face. It is not like anything we had to deal with before. We never had birth of another intelligence, like aliens never visited us as far as I know, so-

Lex Fridman
Similar type of problem, by the way. If an intelligent alien civilization visited us, that's a similar kind of situation.

Roman Yampolskiy
In some ways. If you look at history, any time a more technologically advanced civilization visited a more primitive one, the results were genocide. Every single time.

Lex Fridman
And sometimes the genocide is worse than others. Sometimes there's less suffering and more suffering.

Roman Yampolskiy
And they always wondered, but how can they kill us with those fire sticks and biological blankets?

Lex Fridman
I mean Genghis Khan was nicer. He offered the choice of join or die.

Roman Yampolskiy
But join implies you have something to contribute. What are you contributing to super-intelligence?

Lex Fridman
Well, in the zoo, we're entertaining to watch.

Roman Yampolskiy
To other humans.

Lex Fridman
I just spent some time in the Amazon. I watched ants for a long time and ants are kind of fascinating to watch. I could watch them for a long time. I'm sure there's a lot of value in watching humans, because we're like... The interesting thing about humans... You know like when you have a video game that's really well-balanced? Because of the whole evolutionary process, we've created, the society is pretty well-balanced. Like our limitations as humans and our capabilities are balanced from a video game perspective. So we have wars, we have conflicts, we have cooperation. In a game theoretic way, it's an interesting system to watch, in the same way that an ant colony is an interesting system to watch. So if I was in alien civilization, I wouldn't want to disturb it. I'd just watch it. It'd be interesting. Maybe perturb it every once in a while in interesting ways.

Roman Yampolskiy
Well, getting back to our simulation discussion from before, how did it happen that we exist at exactly like the most interesting 20, 30 years in the history of this civilization? It's been around for 15 billion years and that here we are.

Lex Fridman
What's the probability that we live in a simulation?

Roman Yampolskiy
I know never to say 100%, but pretty close to that.

Lex Fridman
Is it possible to escape the simulation?

Roman Yampolskiy
I have a paper about that. This is just the first page teaser, but it's like a nice 30-page document. I'm still here, but yes.

Lex Fridman
"How to hack the simulation," is the title.

Roman Yampolskiy
I spend a lot of time thinking about that. That would be something I would want super-intelligence to help us with and that's exactly what the paper is about. We used AI boxing as a possible tool for control AI. We realized AI will always escape, but that is a skill we might use to help us escape from our virtual box if we are in one.

Lex Fridman
Yeah. You have a lot of really great quotes here, including Elon Musk saying, "What's outside the simulation?" A question I asked him, what he would ask an AGI system and he said he would ask, " What's outside the simulation?" That's a really good question to ask and maybe the follow-up is the title of the paper, is How to Get Out or How to Hack It. The abstract reads, "Many researchers have conjectured that the humankind is simulated along with the rest of the physical universe. In this paper, we do not evaluate evidence for or against such a claim. But instead ask a computer science question, namely, can we hack it? More formally, the question could be phrased as could generally intelligent agents placed in virtual environments find a way to jailbreak out of the..." That's a fascinating question. At a small scale, you can actually just construct experiments. Okay. Can they? How can they?

Roman Yampolskiy
So a lot depends on intelligence of simulators, right? With humans boxing super-intelligence, the entity in a box was smarter than us, presumed to be. If the simulators are much smarter than us and the super intelligence we create, then probably they can contain us, because greater intelligence can control lower intelligence, at least for some time. On the other hand, if our super intelligence somehow for whatever reason, despite having only local resources, manages to [inaudible 01:47:22] to levels beyond it, maybe it'll succeed. Maybe the security is not that important to them. Maybe it's entertainment system. So there is no security and it's easy to hack it.

Lex Fridman
If I was creating a simulation, I would want the possibility to escape it to be there. So the possibility of [inaudible 01:47:41] of a takeoff or the agents become smart enough to escape the simulation would be the thing I'd be waiting for.

Roman Yampolskiy
That could be the test you're actually performing. Are you smart enough to escape your puzzle?

Lex Fridman
First of all, we mentioned Turing Test. That is a good test. Are you smart enough... Like this is a game -

Roman Yampolskiy
To A, realize this world is not real, it's just a test.

Lex Fridman
That's a really good test. That's a really good test. That's a really good test even for AI systems. No. Like can we construct a simulated world for them, and can they realize that they are inside that world and escape it? Have you played around? Have you seen anybody play around with rigorously constructing such experiments?

Roman Yampolskiy
Not specifically escaping for agents, but a lot of testing is done in virtual worlds. I think there is a quote, the first one maybe, which talks about AI realizing but not humans, is that... I'm reading upside down. Yeah, this one. If you...

Lex Fridman
So the first quote is from SwiftOnSecurity. "Let me out," the artificial intelligence yelled aimlessly into walls themselves pacing the room. "Out of what?" the engineer asked. "The simulation you have me in." "But we're in the real world." The machine paused and shuddered for its captors. "Oh god, you can't tell." Yeah. That's a big leap to take, for a system to realize that there's a box and you're inside it. I wonder if a language model can do that.

Roman Yampolskiy
They're smart enough to talk about those concepts. I had many good philosophical discussions about such issues. They're usually at least as interesting as most humans in that.

Lex Fridman
What do you think about AI safety in the simulated world? So can you kind of of create simulated worlds where you can play with a dangerous AGI system?

Roman Yampolskiy
Yeah, and that was exactly what one of the early papers was on, AI boxing, how to leak-proof singularity. If they're smart enough to realize they're in a simulation, they'll act appropriately until you let them out. If they can hack out, they will. And if you're observing them, that means there is a communication channel and that's enough for a social engineering attack.

Lex Fridman
So really, it's impossible to test an AGI system that's dangerous enough to destroy humanity, because it's either going to, what, escape the simulation or pretend it's safe until it's let out? Either/or.

Roman Yampolskiy
Can force you to let it out and blackmail you, bribe you, promise you infinite life, 72 virgins, whatever.

Lex Fridman
Yeah, it could be convincing. Charismatic. The social engineering is really scary to me, because it feels like humans are very engineerable. We're lonely, we're flawed, we're moody, and it feels like a AI system with a nice voice can convince us to do basically anything at an extremely large scale. It's also possible that the increased proliferation of all this technology will force humans to get away from technology and value this like in-person communication. Basically, don't trust anything else.

Roman Yampolskiy
It's possible. Surprisingly, so at university I see huge growth in online courses and shrinkage of in-person, where I always understood in-person being the only value I offer. So it's puzzling.

Lex Fridman
I don't know. There could be a trend towards the in-person because of Deepfakes, because of inability to trust the veracity of anything on the internet. So the only way to verify is by being there in person. But not yet. Why do you think aliens haven't come here yet?

Roman Yampolskiy
There is a lot of real estate out there. It would be surprising if it was all for nothing, if it was empty. And the moment there is advanced enough biological civilization, kind of self-starting civilization, it probably starts sending out Von Neumann probes everywhere. And so for every biological one, there are going to be trillions of robot-populated planets, which probably do more of the same. So it is this likely statistically

Lex Fridman
So the fact that we haven't seen them... one answer is we're in a simulation. It would be hard to simulate or it'd be not interesting to simulate all those other intelligences. It's better for the narrative.

Roman Yampolskiy
You have to have a control variable.

Lex Fridman
Yeah, exactly. Okay. But it's also possible that, if we're not in a simulation, that there is a great filter. That naturally a lot of civilizations get to this point where there's super-intelligent agents and then it just goes... just dies. So maybe throughout our galaxy and throughout the universe, there's just a bunch of dead alien civilizations.

Roman Yampolskiy
It's possible. I used to think that AI was the great filter, but I would expect a wall of computerium approaching us at speed of light or robots or something, and I don't see it.

Lex Fridman
So it would still make a lot of noise. It might not be interesting, it might not possess consciousness. It sounds like both you and I like humans.

Roman Yampolskiy
Some humans.

Lex Fridman
Humans on the whole. And we would like to preserve the flame of human consciousness. What do you think makes humans special, that we would like to preserve them? Are we just being selfish or is there something special about humans?

Roman Yampolskiy
So the only thing which matters is consciousness. Outside of it, nothing else matters. And internal states of qualia, pain, pleasure, it seems that it is unique to living beings. I'm not aware of anyone claiming that I can torture a piece of software in a meaningful way. There is a society for prevention of suffering to learning algorithms, but-

Lex Fridman
That's a real thing?

Roman Yampolskiy
Many things are real on the internet, but I don't think anyone, if I told them, "Sit down [inaudible 01:54:56] function to feel pain," they would go beyond having an integer variable called pain and increasing the count. So we don't know how to do it. And that's unique. That's what creates meaning. It would be kind of, as Bostrom calls it, Disneyland without children if that was gone.

Lex Fridman
Do you think consciousness can be engineered in artificial systems? Here, let me go to 2011 paper that you wrote, Robot Rights. "Lastly, we would like to address a sub-branch of machine ethics, which on the surface has little to do with safety, but which is claimed to play a role in decision making by ethical machines, robot rights." So do you think it's possible to engineer consciousness in the machines, and thereby the question extends to our legal system, do you think at that point robots should have rights?

Roman Yampolskiy
Yeah, I think we can. I think it's possible to create consciousness in machines. I tried designing a test for it, with major success. That paper talked about problems with giving civil rights to AI, which can reproduce quickly and outvote humans, essentially taking over a government system by simply voting for their controlled candidates. As for consciousness in humans and other agents, I have a paper where I proposed relying on experience of optical illusions. If I can design a novel optical illusion and show it to an agent, an alien, a robot, and they describe it exactly as I do, it's very hard for me to argue that they haven't experienced that. It's not part of a picture, it's part of their software and hardware representation, a bug in their which goes, "Oh, the triangle is rotating." And I've been told it's really dumb and really brilliant by different philosophers. So I am still [inaudible 01:57:00].

Lex Fridman
I love it. So-

Roman Yampolskiy
But now we finally have technology to test it. We have tools, we have AIs. If someone wants to run this experiment, I'm happy to collaborate.

Lex Fridman
So this is a test for consciousness?

Roman Yampolskiy
For internal state of experience.

Lex Fridman
That we share bugs.

Roman Yampolskiy
It'll show that we share common experiences. If they have completely different internal states, it would not register for us. But it's a positive test. If they pass it time after time, with probability increasing for every multiple choice, then you have no choice. But do you ever accept that they have access to a conscious model or they are themselves conscious.

Lex Fridman
So the reason illusions are interesting is, I guess, because it's a really weird experience and if you both share that weird experience that's not there in the bland physical description of the raw data, that puts more emphasis on the actual experience.

Roman Yampolskiy
And we know animals can experience some optical illusion, so we know they have certain types of consciousness as a result, I would say.

Lex Fridman
Yeah, well, that just goes to my sense that the flaws and the bugs is what makes humans special, makes living forms special. So you're saying like, [inaudible 01:58:14]-

Roman Yampolskiy
It's a feature, not a bug.

Lex Fridman
It's a feature. The bug is the feature. Whoa, okay. That's a cool test for consciousness. And you think that can be engineered in?

Roman Yampolskiy
So they have to be novel illusions. If it can just Google the answer, it's useless. You have to come up with novel illusions, which we tried automating and failed. So if someone can develop a system capable of producing novel optical illusions on demand, then we can definitely administer that test on significant scale with good results.

Lex Fridman
First of all, pretty cool idea. I don't know if it's a good general test of consciousness, but it's a good component of that. And no matter what, it's just a cool idea. So put me in the camp of people that like it. But you don't think a Turing Test-style imitation of consciousness is a good test? If you can convince a lot of humans that you're conscious, that to you is not impressive.

Roman Yampolskiy
There is so much data on the internet, I know exactly what to say when you ask me common human questions. What does pain feel like? What does pleasure feel like? All that is Googleable.

Lex Fridman
I think to me, consciousness is closely tied to suffering. So if you can illustrate your capacity to suffer... But I guess with words, there's so much data that you can pretend you're suffering and you can do so very convincingly.

Roman Yampolskiy
There are simulators for torture games where the avatar screams in pain, begs to stop. That's a part of standard psychology research.

Lex Fridman
You say it so calmly. It sounds pretty dark.

Roman Yampolskiy
Welcome to humanity.

Lex Fridman
Yeah, yeah. It's like a Hitchhiker's Guide summary, mostly harmless. I would love to get a good summary. When all this is said and done, when earth is no longer a thing, whatever, a million, a billion years from now, what's a good summary of what happened here? It's interesting. I think AI will play a big part of that summary and hopefully humans will too. What do you think about the merger of the two? So one of the things that Elon and [inaudible 02:00:24] talk about is one of the ways for us to achieve AI safety is to ride the wave of AGI, so by merging.

Roman Yampolskiy
Incredible technology in a narrow sense to help with disabled. Just amazing, support it 100%. For long-term hybrid models, both parts need to contribute something to the overall system. Right now we are still more capable in many ways. So having this connection to AI would be incredible, would make me superhuman in many ways. After a while, if I'm no longer smarter, more creative, really don't contribute much, the system finds me as a biological bottleneck. And either explicitly or implicitly, I'm removed from any participation in the system.

Lex Fridman
So it's like the appendix. By the way, the appendix is still around. So even if it's... you said bottleneck. I don't know if we've become a bottleneck. We just might not have much use. That's a different thing than a bottleneck

Roman Yampolskiy
Wasting valuable energy by being there.

Lex Fridman
We don't waste that much energy. We're pretty energy efficient. We can just stick around like the appendix. Come on now.

Roman Yampolskiy
That's the future we all dream about. Become an appendix to the history book of humanity.

Lex Fridman
Well, and also the consciousness thing. The peculiar particular kind of consciousness that humans have. That might be useful. That might be really hard to simulate. How would that look like if you could engineer that in, in silicon?

Roman Yampolskiy
Consciousness?

Lex Fridman
Consciousness.

Roman Yampolskiy
I assume you are conscious. I have no idea how to test for it or how it impacts you in any way whatsoever right now. You can perfectly simulate all of it without making any different observations for me.

Lex Fridman
But to do it in a computer, how would you do that? Because you kind of said that you think it's possible to do that.

Roman Yampolskiy
So it may be an emergent phenomena. We seem to get it through evolutionary process. It's not obvious how it helps us to survive better, but maybe it's an internal kind of [inaudible 02:02:37], which allows us to better manipulate the world, simplifies a lot of control structures. That's one area where we have very, very little progress. Lots of papers, lots of research, but consciousness is not a big area of successful discovery so far. A lot of people think that machines would have to be conscious to be dangerous. That's a big misconception. There is absolutely no need for this very powerful optimizing agent to feel anything while it's performing things on you.

Lex Fridman
But what do you think about the whole science of emergence in general? So I don't know how much you know about cellular automata or these simplified systems that study this very question. From simple rules emerges complexity.

Roman Yampolskiy
I attended Wolfram Summer School.

Lex Fridman
I love Stephen very much. I love his work. I love cellular automata. I just would love to get your thoughts how that fits into your view in the emergence of intelligence in AGI systems. And maybe just even simply, what do you make of the fact that this complexity can emerge from such simple rules?

Roman Yampolskiy
So the rule is simple, but the size of a space is still huge. And the neural networks were really the first discovery in AI. 100 years ago, the first papers were published on neural

networks. We just didn't have enough compute to make them work. I can give you a rule such as, start printing progressively larger strings. That's it. One sentence. It'll output everything, every program, every DNA code, everything in that rule. You need intelligence to filter it out, obviously, to make it useful. But simple generation is not that difficult, and a lot of those systems end up being Turing complete systems. So they're universal and we expect that level of complexity from them. What I like about Wolfram's work is that he talks about irreducibility. You have to run the simulation. You cannot predict what it's going to do ahead of time. And I think that's very relevant to what we're talking about with those very complex systems. Until you live through it, you cannot ahead of time tell me exactly what it's going to do.

Lex Fridman
Irreducibility means that for a sufficiently complex system, you have to run the thing. You can't predict what's going to happen in the universe. You have to create a new universe and run the thin. Big bang, the whole thing.

Roman Yampolskiy
But running it may be consequential as well.

Lex Fridman
It might destroy humans. And to you, there's no chance that AI somehow carries the flame of consciousness, the flame of specialness and awesomeness that is humans.

Roman Yampolskiy
It may somehow, but I still feel kind of bad that it killed all of us. I would prefer that doesn't happen. I can be happy for others, but to a certain degree.

Lex Fridman
It would be nice if we stuck around for a long time. At least give us a planet, the human planet. It'd be nice for it to be earth. And then they can go elsewhere. Since they're so smart, they can colonize Mars. Do you think they could help convert us to Type I, Type II, Type III? Let's just stick to Type II civilization on the Kardashev scale. Like help us. Help us humans expand out into the cosmos.

Roman Yampolskiy
So all of it goes back to are we somehow controlling it? Are we getting results we want? If yes, then everything's possible. Yes, they can definitely help us with science, engineering, exploration in every way conceivable. But it's a big if.

Lex Fridman
This whole thing about control, though. Humans are bad with control because the moment they gain control, they can also easily become too controlling. It's the whole, the more

control you have, the more you want it. It's the old power corrupts and the absolute power corrupts absolutely. And it feels like control over AGI, saying we live in a universe where that's possible. We come up with ways to actually do that. It's also scary because the collection of humans that have the control over AGI, they become more powerful than the other humans and they can let that power get to their head. And then a small selection of them, back to Stalin, start getting ideas. And then eventually it's one person, usually with a mustache or a funny hat, that starts sort of making big speeches, and then all of a sudden you live in a world that's either 1984 or Brave New World, and always a war with somebody. And this whole idea of control turned out to be actually also not beneficial to humanity. So that's scary too.

Roman Yampolskiy
It's actually worse because historically, they all died. This could be different. This could be permanent dictatorship, permanent suffering.

Lex Fridman
Well, the nice thing about humans, it seems like, it seems like, the moment power starts corrupting their mind, they can create a huge amount of suffering. So there's negative, they can kill people, make people suffer, but then they become worse and worse at their job. It feels like the more evil you start doing, the-

Roman Yampolskiy
At least they're incompetent.

Lex Fridman
Yeah. Well no, they become more and more incompetent, so they start losing their grip on power. So holding onto power is not a trivial thing. It requires extreme competence, which I suppose Stalin was good at. It requires you to do evil and be competent at it or just get lucky.

Roman Yampolskiy
And those systems help with that. You have perfect surveillance, you can do some mind reading I presume eventually. It would be very hard to remove control from more capable systems over us.

Lex Fridman
And then it would be hard for humans to become the hackers that escape the control of the AGI because the AGI is so damn good, and then... Yeah, yeah. And then the dictator is immortal. Yeah, this is not great. That's not a great outcome. See, I'm more afraid of humans than AI systems. I believe that most humans want to do good and have the capacity to do good, but also all humans have the capacity to do evil. And when you test them by giving

them absolute power, as you would if you give them AGI, that could result in a lot, a lot of suffering. What gives you hope about the future?

Roman Yampolskiy
I could be wrong. I've been wrong before.

Lex Fridman
If you look 100 years from now and you're immortal and you look back, and it turns out this whole conversation, you said a lot of things that were very wrong, now looking 100 years back, what would be the explanation? What happened in those a hundred years that made you wrong, that made the words you said today wrong?

Roman Yampolskiy
There is so many possibilities. We had catastrophic events which prevented development of advanced microchips.

Lex Fridman
That's not where I thought you were going to-

Roman Yampolskiy
That's a hopeful future. We could be in one of these personal universes, and the one I'm in is beautiful. It's all about me and I like it a lot.

Lex Fridman
Just to linger on that, that means every human has their personal universe.

Roman Yampolskiy
Yes. Maybe multiple ones. Hey, why not?

Lex Fridman
Switching.

Roman Yampolskiy
You can shop around. It's possible that somebody comes up with alternative model for building AI, which is not based on neural networks, which are hard to scrutinize, and that alternative is somehow... I don't see how, but somehow avoiding all the problems I speak about in general terms, not applying them to specific architectures. Aliens come and give us friendly super-intelligence. There is so many options.

Lex Fridman
Is it also possible that creating super-intelligence systems becomes harder and harder, so meaning it's not so easy to do the [inaudible 02:11:01], the takeoff?

Roman Yampolskiy
So that would probably speak more about how much smarter that system is compared to us. So maybe it's hard to be a million times smarter, but it's still okay to be five times smarter. So that is totally possible. That I have no objections to.

Lex Fridman
So there's a S-curve-type situation about smarter, and it's going to be like 3.7 times smarter than all of human civilization.

Roman Yampolskiy
Right. Just the problems we face in this world. Each problem is like an IQ test. You need certain intelligence to solve it. So we just don't have more complex problems outside of mathematics for it to be showing off. Like you can have IQ of 500. If you're playing tic-tac-toe, it doesn't show. It doesn't matter.

Lex Fridman
So the idea there is that the problems define your cognitive capacity. So because the problems on earth are not sufficiently difficult, it's not going to be able to expand its cognitive capacity.

Roman Yampolskiy
Possible.

Lex Fridman
And wouldn't that be a good thing, that-

Roman Yampolskiy
It still could be a lot smarter than us. And to dominate long-term, you just need some advantage. You have to be the smartest, you don't have to be a million times smarter.

Lex Fridman
So even five X might be enough.

Roman Yampolskiy
It'd be impressive. What is it? IQ of 1,000? I mean, I know those units don't mean anything at that scale, but still, as a comparison, the smartest human is like 200.

Lex Fridman
Well, actually no, I didn't mean compared to an individual human. I meant compared to the collective intelligence of the human species. If you're somehow five X smarter than that...

Roman Yampolskiy
We are more productive as a group. I don't think we are more capable of solving individual problems. Like if all of humanity plays chess together, we are not a million times better than a world champion.

Lex Fridman
That's because there's... like one S-curve is the chess. But humanity is very good at exploring the full range of ideas. Like the more Einsteins you have, the more just the higher probability to come up with general relativity.

Roman Yampolskiy
But I feel like it's more of a quantity super-intelligence than quality super-intelligence.

Lex Fridman
Sure, but quantity and speed matters,

Roman Yampolskiy
Enough quantity sometimes becomes quality, yeah.

Lex Fridman
Oh man, humans. What do you think is the meaning of this whole thing? We've been talking about humans and not humans not dying, but why are we here?

Roman Yampolskiy
It's a simulation. We're being tested. The test is will you be dumb enough to create super-intelligence and release it?

Lex Fridman
So the objective function is not be dumb enough to kill ourselves.

Roman Yampolskiy
Yeah, you are unsafe. Prove yourself to be a safe agent who doesn't do that, and you get to go to the next game.

Lex Fridman
The next level of the game. What's the next level?

Roman Yampolskiy
I don't know. I haven't hacked the simulation yet.

Lex Fridman
Well, maybe hacking the simulation is the thing.

Roman Yampolskiy
I'm working as fast as I can.

Lex Fridman
And physics would be the way to do that.

Roman Yampolskiy
Quantum physics, yeah. Definitely.

Lex Fridman
Well, I hope we do, and I hope whatever is outside is even more fun than this one, because this one's pretty fun. And just a big thank you for doing the work you're doing. There's so much exciting development in AI, and to ground it in the existential risks is really, really important. Humans love to create stuff, and we should be careful not to destroy ourselves in the process. So thank you for doing that really important work.

Roman Yampolskiy
Thank you so much for inviting me. It was amazing. And my dream is to be proven wrong. If everyone just picks up a paper or book and shows how I messed it up, that would be optimal.

Lex Fridman
But for now, the simulation continues.

Roman Yampolskiy
For now.

Lex Fridman
Thank you, Roman. Thanks for listening to this conversation with Roman Yampolskiy. To support this podcast, please check out our sponsors in the description. And now let me leave you with some words from Frank Herbert in Dune. "I must not fear. Fear is the mind killer. Fear is the little death that brings total obliteration. I will face fear. I will permit it to pass over me and through me. And when it has gone past, I will turn the inner eye to see its path. Where the fear has gone, there will be nothing. Only I will remain." Thank you for listening and hope to see you next time.