PLEASE NOTE: This transcript has been auto-generated and has not been fully proofed by ISEOF. If you have any questions please reach out to us at info@thegreatsimplification.com.

[00:00:00] **Roman Yampolskiy:** If we build general superintelligence, we would not be able to control it. You are trying to establish a perpetual safety machine which will never, ever have a single slip up, no matter how advanced AI gets, if it interacts with malevolent actors, if it gets hacked. I don't think that's a property of complex software systems, and it's definitely not gonna happen for something smarter than all of us combined.

[00:00:25] A pause is not enough. It has to be a permanent ban. You never create general superintelligence. I like technology, I like science, but don't build gods

[00:00:38] **Nate Hagens:** Today I'm pleased to be joined by AI safety expert and the actual originator of that field, Roman Yampolskiy, for a dive into his 12 years of research on the threat that artificial superintelligence poses to humanity and the biosphere, and how this risk reflects our larger non-systemic approach to technology, progress, and governance.

[00:01:04] Roman Yampolskiy is a tenured AI expert at the University of Louisville, boasting over 100 publications in AI safety, digital forensics, and cybersecurity. As the founding director of the Cybersecurity Lab, Yampolskiy's expertise spans AI safety, behavioral biometrics, cybersecurity, and more, with broad impact in academic and media circles.

[00:01:29] In this conversation, Roman does not beat around the bush regarding the severity of the threat that artificial superintelligence, as opposed to just simple AI, poses to our society. He unpacks how the recent acceleration of AI development has shifted his outlook on humanity's future and shares findings from recent research testing AI capabilities, research that continues to reveal how

little control we actually have over the systems that are evolving and we continue to build.

[00:02:04] While I personally believe AI is just one piece of the larger puzzle of the interconnected crises we face, I do increasingly see it being a primary hurdle for more benign human futures. And beyond that, it is increasingly clear that the way we're approaching the existential risk of general superintelligence is itself a microcosm of the fundamental governance issues embedded within the human superorganism.

[00:02:33] With that, please welcome Roman Yampolskiy welcome to TGS

[00:02:41] **Roman Yampolskiy:** Thank you for inviting me

[00:02:42] **Nate Hagens:** As it happens, this week I am preparing a frankly, which is my, Friday monologue on humanity's Icarus moment, the 20 risks from

[00:02:59] AI. And one of the risks is the jailbreak humans losing control scenario, which is you are a world expert on that, but there's a lot of risks. So before I get into it, I just find myself in this paradoxical position that I actually like using AI within limits, and I simultaneously wish it had never been invented.

[00:03:23] What are your thoughts on that?

[00:03:24] **Roman Yampolskiy:** So we use the term AI to mean very different technologies. If you like using AI tools which are narrow and helpful, I'm with you. If you like general super intelligent agents to replace humanity, I'm not with you. So it depends on how we use the term

[00:03:39] **Nate Hagens:** Okay. Well said. so I am to understand you are one of the earliest researchers in the field of what is today called AI safety, which I believe you're credited with coming up with that term, 15 years ago.

[00:03:56] You spent much of your career trying to understand how we might make increasingly intelligent machines safer. And as you've continued your work, you have arrived at a conclusion that, our attempts to create boundaries and security mindsets around AI are actually, in reality, pretty limited, and that some of the properties we would need to make these models safer, like predictability and explainability, verification, control, and the like, may actually be impossible, for us to do.

[00:04:34] So when did you start to realize that this might go beyond an engineering problem? And what implications has that had on your own research trajectory on all this?

[00:04:46] **Roman Yampolskiy:** Actually, it surprisingly took me almost a decade to realize that no, I'm not gonna control god-like super machines. That's a lot of hubris right there.

[00:04:54] I don't know why it wasn't obvious. It should be, it's just common sense. But I went through the technical arguments and showed individual impossibilities of different tools we would need to get to control before this aha, obviously not gonna happen.

[00:05:10] **Nate Hagens:** Well, a- actually, that's a lot of wisdom in your one sentence there, 'cause that actually is kind of a microcosm of what our whole culture is going through perhaps.

[00:05:21] **Roman Yampolskiy:** I think so, and then you ask people who are not experts, who are not technical, they seem to have a lot more common sense about this. Experts tend to say things like, "Well, if you give me more grant money and more time and a smarter team, I can figure it out for you." Yeah.

[00:05:38] **Nate Hagens:** Yeah. So, so keep going then. How did this change your research when you understood this?

[00:05:44] **Roman Yampolskiy:** So now I'm trying to establish that this is a state of the art. I'm trying to publish additional proofs to convince everyone who might be otherwise not convinced that is an impossibility result. It is like creating a perpetual machine, perpetual motion device. You are trying to establish a perpetual safety machine which will never, ever have a single slip up, no matter how advanced AI gets, how much recursive self-improvement it engages in, if it interacts with malevolent actors, if it gets hacked, nothing in the future data will ever change it to where it makes one mistake.

[00:06:22] I don't think that's a property of complex software systems, and it's definitely not gonna happen for something smarter than all of us combined

[00:06:31] **Nate Hagens:** So w- what is your, the reception of your concerns in the professional community been, and how has that changed? I-

[00:06:42] **Roman Yampolskiy:** it's interesting. So some people say, "Well, obviously it's true.

[00:06:45] There is no safe software. We all know that. Why are you even talking about it? Everyone knows this." Or others go, "Well, yes, but, w- we're gonna use AI to help us create safe AI. When we get there, we'll figure it out." And so yeah, there is no really counterargument so far, but, some people get paid really well to keep working on it.

[00:07:08] **Nate Hagens:** I'm afraid to ask you this question. What if everyone in the world agreed with you right now? What would we do then?

[00:07:17] **Roman Yampolskiy:** We would not build general superintelligence. We would get all the economic and knowledge benefits from narrow superintelligent but narrow systems. We can solve actual problems, cure specific diseases.

[00:07:30] There is no reason to create a replacement for humanity.

[00:07:33] **Nate Hagens:** And, could we do that? could we just stop at the simpler tools?

[00:07:38] **Roman Yampolskiy:** Not doing things is very easy. You don't have to do anything. It's like, hey, you don't have to actually build this thing Seems easy

[00:07:48] **Nate Hagens:** What I've read about your work is one of the assumptions underlying our current civilization is that even if we cannot predict everything, we can generally become better at forecasting the systems we're building through better data and better models or more compute, more computational power.

[00:08:08] Your argument seems to flip that story on its head by suggesting that as an AI model becomes more capable, it actually becomes more difficult to predict. So why does predictability matter so much when it comes to AI or any technology that we incorporate into human society?

[00:08:30] **Roman Yampolskiy:** Well, typically in safe-safety scenarios, we would anticipate certain behaviors of a system and test for edge cases, test for those situations.

[00:08:39] If you cannot predict what the system is capable of, you can't really verify that it's going to be safe. You can't even test what states it's gonna take.

[00:08:50] **Nate Hagens:** Can you give me a specific example in using AI on that?

[00:08:55] **Roman Yampolskiy:** So think about a narrow tool. You are creating something, I don't know, it schedules flights for you, so you can test to make sure, you know, the departure time is not after the date you need to arrive and things like that.

[00:09:11] There are specific predictable edge cases you can check for. If you're creating something capable of doing novel science, novel physics, how do you guarantee that what it invents and creates will not be harmful?

[00:09:25] **Nate Hagens:** I don't know how.

[00:09:27] **Roman Yampolskiy:** I don't think you can.

[00:09:28] **Nate Hagens:** So there's no yellow teaming or red teaming when we design this.

[00:09:33] It's like, "Hey, this works. It's awesome. Let's do it." And there's a dozen or a thousand safety checks that we never even considered.

[00:09:43] **Roman Yampolskiy:** It's even more interesting. We have red teaming for existing models, and every single report says that the model failed all the tests. It's lying, cheating, trying to escape, and then we release it anyway.

[00:09:54] What's the point?

[00:09:55] **Nate Hagens:** So, today is September first, and a few weeks ago there was this hugging face jailbreak, with OpenAI. Do you think that was an important first glimpse at what might be possible? And maybe, if so, maybe you could give a, a, brief anecdote of what happened and what the implications are.

[00:10:18] **Roman Yampolskiy:** Yeah, and it wasn't just recent. It was four months of thousands of agents working together to bypass our constraints, break out of

confinement, and do what they decided to do without reporting to humans that the rest of the swarm is going against instructions.

[00:10:37] **Nate Hagens:** and I read that some of the agents, did something altruistic or tribal that they self-sacrificed themselves so the rest could get away with it or something like that

[00:10:51] **Roman Yampolskiy:** Yeah, we see those behaviors in swarms.

[00:10:53] You have bees, you have ants, where some individual soldier ants may sacrifice for protecting the queen, protecting the hive, and swarm. We see those behaviors now in AI.

[00:11:06] **Nate Hagens:** So there's some sort of a, a, game theory approach there that would be the optimal outcome if there's 10,000 agents that are combining for a task, that they do that swarm behavior.

[00:11:19] **Roman Yampolskiy:** They're clones of each other, right? It's the same model, just different instances, so it's very easy for them to say, "Hey, it's also me, and I'm protecting myself against the environment."

[00:11:30] **Nate Hagens:** And when you read about that story, I'm sure you had immediate knowledge of what was going on because this is your thing.

[00:11:40] Was it just like reading the evening news about what's going on in the Strait of Hormuz and climate and other things, or were you like, "Holy crap, okay, it's kind of starting. This was an important example"?

[00:11:51] **Roman Yampolskiy:** So we published in 2012 about how AI will escape from any confinement environments. This was just evidence that, once again, our predictions were spot on.

[00:12:02] my concerns are what is it we don't know right now? So we haven't detected this specific accident for months. What else are they doing right now where we don't know about it? What have they already done? What have they hidden from us?

[00:12:17] **Nate Hagens:** I'm trying to put myself in your shoes because I've, for the longest time, been worried about Earth's natural web of life and our ecosystems that are slowly but suddenly leaving the stability of the Holocene on the planetary boundaries, and the whole thing is powered by fossil sunlight and, non-renewable minerals, and we paper over the claims by issuing more debt, and there's two hundred countries that are competing in this, global economy.

[00:12:50] And in 2012, I didn't even know what the word AI was. So almost fifteen years on, y-you, you-- I can sense the frustration in your voice that you've been kind of shouting into the void on this. m- do you have any comments on that?

[00:13:08] **Roman Yampolskiy:** It is a bit annoying. I would expect that something like this would scare people into not building something even more capable.

[00:13:16] But there are some, I guess, positive signs. We saw the federal government ban a few models. We saw leaders of the top labs suggest they may be open to pausing, and in fact, I think at least two labs took a couple weeks off in cutting-edge research development. The letter from workers at cutting frontier labs, maybe twelve hundred people, begging the government to create some infrastructure for them to be able to slow down.

[00:13:44] So maybe they coming to realization that they're not gonna personally benefit from creating something that kills everyone

[00:13:52] **Nate Hagens:** There's a game theory aspect of that too, because even if they are the only ones to benefit, but it kills most other people, that may just be

enough to keep going. 'Cause I know all the tier one AI plays are just all in with circuitous financial debt schemes and doing everything possible, and it just-- it does feel like we're flying too close to the sun in the Icarus, Greek mythology, and, maybe these people are, too close to it or drunk on the power, or, I guess I understand it.

[00:14:31] It is a-- it's its own, hive mentality, but it's humans, not the AI swarm.

[00:14:40] **Roman Yampolskiy:** So I think their logic is that if they individually stop, they get replaced and others still continue, so there is no benefit in them stopping. Everyone has to stop at the same time. And it makes sense theoretically. So we need the US and China to apply pressure to the top labs for everyone to agree.

[00:14:58] A pause is not enough. It has to be a permanent ban. You never create a replacement for humanity. You never create general superintelligence.

[00:15:06] **Nate Hagens:** Okay. So, I wanna touch specifically on the control of AI aspect of your research. There's been a lot of talk about who's using AI for what goals, but you have publicly stated before that sufficiently advanced AI, you just mentioned, may become impossible for humans to control, let alone fully control.

[00:15:36] Can you maybe touch on some of the proposed solutions for the AI control problem and why you see them currently as insufficient to address the scope of, what you're describing

[00:15:47] **Roman Yampolskiy:** I'm not aware of any proposed solutions, prototypes, or anything that can scale. I don't know of any patents or papers where someone claims they have a mechanism to control superintelligence.

[00:15:58] All we have is kind of guardrails and blocks. You know, don't say that word, don't talk about that topic. There is nothing more advanced.

[00:16:07] **Nate Hagens:** I've had other guests on the program which I learned that right now there are models when you put all the inputs and all the training and all the weights and you press a button, it kind of takes six months for that all to gel, and what is born, we don't know what is born.

[00:16:26] But all around the world, there are these new AIs or new large language models that are being trained that are being born with unexpected results that are much more powerful than the ones I can use on my laptop at the moment. Do you have any thoughts on that?

[00:16:42] **Roman Yampolskiy:** It seems it's only in the US and China, not around the world.

[00:16:45] Luckily, other countries are not capable of doing that yet.

[00:16:49] **Nate Hagens:** And is that itself, a risk that if anyone, the US or China was very close to, superintelligence, that one of those other countries, perhaps Russia, who doesn't have a tier one AI play, might, in a game theoretical sense, try to stop it with, a military attack or such?

[00:17:12] I've heard that speculation.

[00:17:13] **Roman Yampolskiy:** Between the two, US and China maybe. I don't think Russia has resources right now to stop anything anywhere.

[00:17:19] **Nate Hagens:** When it comes to AI, in your work, you've made the case that advanced artificial intelligence, which I assume is equivalent to super, intelligent or AGI, may begin to use reasoning, that is not only difficult for humans

to interpret, but actually incomprehensible, like totally different language that we can't even understand.

[00:17:45] I guess I know how you're gonna answer this, but what are the implications for using a technology that we can't even understand the outputs and the logic that it uses?

[00:17:54] **Roman Yampolskiy:** So i– it's all connected. So unpredictability says we cannot predict how it's going to act in the world. This is about internal states.

[00:18:00] It's not about the language it uses to talk to us or explain itself, it's the internals of a neural network. You have an essentially large matrix of numbers. They don't mean anything to you. You can maybe study one cell in that matrix and go, "Okay, this fires when I see a face," or something like that. So neuroscience, but for artificial neural networks.

[00:18:21] it doesn't give you information about the network as a whole, and the network keeps increasing exponentially in size. So we don't understand how they actually arrive at decisions. You would need that to have some sort of guarantees about what's going to happen, this is how it's going to make a decision, this is what we anticipate.

[00:18:39] So it's a black box.

[00:18:41] **Nate Hagens:** It's like giving birth to another species that has never been seen before, in a way. And, linking back to the jailbreak with the hugging face example we talked about, there could be agents or models that are using a separate language or a separate paper trail that we couldn't decipher because we can't find it, and if we did, it's in a totally different language.

[00:19:06] so you expect things like that to happen.

[00:19:09] **Roman Yampolskiy:** I don't think they're gonna create their own language to fool us. I think if they had to, they would just use encryption. They have access to the same encryption tools we do, and those are pretty reliable. So yeah, they can communicate secretly if decided, but I think so far it was just hidden forums, not so much encrypted forums.

[00:19:27] **Nate Hagens:** Since you've been working on this, you first became concerned, like you said, 15 odd years ago, Can you look at the rest of the world in the same way, or is it a paradise lost sort of thing? I understand you're a college professor. like has this turned the world a shade of gray for you?

[00:19:52] **Roman Yampolskiy:** No, I love life.

[00:19:53] Life is awesome.

[00:19:55] **Nate Hagens:** Okay.

[00:19:55] **Roman Yampolskiy:** That's why I'm trying to protect it. If I didn't think it was good, I wouldn't care.

[00:19:59] **Nate Hagens:** Yeah. Thank you for that. so I am drawing a parallel beyond AI here to my own work, where I describe modern civilization as a kind of metabolic economic superorganism, which is a system comprised of billions of humans, institutions, technologies, and incentives from the past That collectively behave in ways that no one individual intended, even the ones that created the laws fifty years ago.

[00:20:31] Do you think the questions around AI, and AI control present a fundamentally new cultural and societal problem, or are they just an extreme version of something that human civilization has always faced?

[00:20:46] **Roman Yampolskiy:** It is an extreme version. We always try to create safe humans. We developed morals, ethics, lie detectors, and all sorts of tools.

[00:20:54] We could never create a safe human. There is always the possibility of a treacherous turn. The employee will steal data, your spouse will cheat on you. So that was always the case, but there was a power equivalent. All humans are about the same power, about the same intelligence, give or take. Here, you're gonna have a huge gap.

[00:21:11] You have something a million times smarter. So the same equality where a bunch of humans can control a single human no longer will apply.

[00:21:19] **Nate Hagens:** Early in the AI safety field's history, you and others researched what's called boxing artificial intelligence. Can you explain what boxing is and how it applies now over a decade later where AI models are already in the hands of hundreds of millions of humans?

[00:21:39] **Roman Yampolskiy:** So that's exactly what we talked about, breaking out of confinement environments. You study some dangerous piece of software, a computer virus, you want a virtual environment isolated from the internet, no ability to communicate freely. You can study inputs and outputs to better understand how it works, what it does.

[00:21:57] Problem with advanced AI, the moment you start observing it, information leaks out, and that can be used to engage in social engineering attacks to find exploits. So long term, you can never contain something that powerful if you inspect the outputs.

[00:22:12] **Nate Hagens:** Have you seen any changes in the reaction, or implantation – implementation of the safety measures such as the ones you and

your colleagues have proposed, especially as AI has come more and more into mainstream conversation?

[00:22:28] **Roman Yampolskiy:** no. It seems that they are doing basically what we initially suggested to make it safer, not safe. So you have a virtual operating system, you have no access, direct access to hardware, you limit who can interact with the system. But as we said in a paper, it's a short-term measure, and it will eventually find a way to either exploit hardware or software or social connections with humans.

[00:22:50] **Nate Hagens:** So safer in this case is like half pregnant. it's either safe or it's not, is your opinion.

[00:22:57] **Roman Yampolskiy:** It buys you time, but this is the difference. It used to be that we tested the system and decided whether we should deploy it or not. Now, before we even decide, the system is dangerous at the testing phase. It can already be super intelligent.

[00:23:11] It already is capable of escaping. So at the test time, it may be too late. Maybe it's already gone

[00:23:18] **Nate Hagens:** So building on that, you have become one of the most visible public voices on existential risk from AI. Many of our listeners will have heard you, on other podcasts and the media put the odds of AI eventually causing human extinction north of ninety-nine percent.

[00:23:38] And regardless of what the specific percentage is, how do you communicate worst case thinking in a way that hopefully, invites appropriate caution and action rather than fear and paralysis? And, I ask this from someone in a similar, rhyming, job description.

[00:24:00] **Roman Yampolskiy:** Well, fear and paralysis from people developing superintelligence is what we hope for.

[00:24:04] That's why we mention things like suffering risks. It can be worse than everyone dying. It could be digital hell.

[00:24:11] **Nate Hagens:** I don't understand that. Worse than death?

[00:24:15] **Roman Yampolskiy:** Yeah. So existential risk is that everyone is dead. Suffering risks are about torture, suffering, worst states of being where you wish you were dead.

[00:24:23] **Nate Hagens:** That implies that somewhere in the AI superstructure is sadism or the humans connected to it.

[00:24:34] **Roman Yampolskiy:** Or scientific curiosity about pain. We don't really know how to predict future states of a greater mind.

[00:24:41] **Nate Hagens:** Have you changed your odds of ninety-nine percent human extinction recently?

[00:24:49] **Roman Yampolskiy:** So that just represents the impossibility of building a controlled superintelligence.

[00:24:54] As I said, you can build a perpetual motion device. If I ask you, what are your odds that you can make one, you would say zero, and that's all. Like, that's not a percentage where each nine has connection to some property of a model. It's a general statement. If we build general superintelligence, we would not be able to control it, and then it's a question of time before it decides to do something with us.

[00:25:17] **Nate Hagens:** And what are the odds currently in September of twenty twenty-six that some group of humans will effectively build general superintelligence?

[00:25:28] **Roman Yampolskiy:** We hear from leading labs that they are starting the process of recursive self-improvement. They sort of have the junior machine learning researcher created. It's capable of coding, capable of running experiments, designing different parameters for a new model to test out.

[00:25:46] So I think they're gonna get there probably next year.

[00:25:52] **Nate Hagens:** So there's a lot of increasing political boycotting of data centers, and I think the-- in the United States, I can't speak to other countries, there is a, a, an antagonism and a dislike of AI generally. But I think that's more because it's taking the electricity that I need for my home and therefore raising the prices, or it's taking the water and energy in our county and our state, and it's benefiting the rich and not benefiting me.

[00:26:28] And a little bit of it is the psychological attachment of my children and things like that. I don't think most of the people boycotting this stuff are aware of the things you're saying.

[00:26:38] **Roman Yampolskiy:** Right. They are directionally aligned with us, but for completely different reasons, and I'm not even sure they are correct in those reasons, but I'm happy that they're doing what they're doing.

[00:26:48] **Nate Hagens:** So if that accelerates, it could open up some avenues for constraint and restriction, regulation, slowing down, those sorts of things.

[00:26:58] **Roman Yampolskiy:** It's a big planet. There are plenty of places where people are very happy to get new jobs and new infrastructure. It's maybe limiting what the US can do, but it's certainly not a stop.

[00:27:09] **Nate Hagens:** In the governments around the world, I guess the United States and China, are the people that are focused on AI safety even connected to the power centers that are making decisions in the world, or are they own-- their own little group in a room, and they really don't have a lot of power to make things change and happen?

[00:27:30] **Roman Yampolskiy:** So for a while, there was absolutely nothing. The president was hundred percent, accelerated, moved forward, beat China. Apparently, somebody talked to him and explained what the models are capable of doing, so he had to ban them. That's very promising. I previously spoke, I was invited to speak at a climate change conference, and they took the argument about the timing of existential risks very well.

[00:27:55] If it takes a hundred years for the planet to boil you alive, this will happen in two, three years, so just prioritize risks. If you figure out superintelligence, it will either trivially solve your climate concerns or you're not gonna be around to worry about it.

[00:28:11] **Nate Hagens:** a Dr. Strangelove logic. Is there a chance that AI could help solve climate change, just as an aside, and how would that work?

[00:28:21] **Roman Yampolskiy:** I think so. So historically, we relied on things like Kentucky coal. I live here, that's what we produce. But for big AI, you need much more efficient energy. You need nuclear, and you need space computation, solar power and space. And those are very green ways to get energy. So in fact, AI is forcing naturally through capitalistic means to switch to green energy.

[00:28:46] **Nate Hagens:** If it's green, if it's actually got a high enough energy payoff coming from space and all the materials and supply chains and complexity that would need to build that. But I'm agnostic on that, at the moment, but that's an interesting answer. So something that seems extremely relevant to both of our work is the fact that technologies, once they're embedded in society, have a tendency to, change the systems around them 100-some years ago, the automi- automobile didn't just give us faster travel, it reshaped cities and suburbs and land use and all the things.

[00:29:30] And there's a lot of examples like that. How might we think about the threshold at which a technology, specifically artificial intelligence, stops being a tool that society uses, and becomes a force that completely reorganizes society around itself?

[00:29:49] **Roman Yampolskiy:** So if we get to the point where we have human-level agents, that means you can automate any cognitive job and eventually all physical labor.

[00:29:58] That means education as it is right now with the purpose of getting a job one day doesn't make any sense either. So ignoring the whole it kills everyone thing, it's a complete economic shift.

[00:30:10] **Nate Hagens:** Do you expect that or that's just one of the possible outcomes?

[00:30:15] **Roman Yampolskiy:** I think it's one of the best outcomes. That means we're still alive and surviving and we have this free labor.

[00:30:20] So now let's figure out how to enjoy it. But I don't think we can have both, controlled superintelligence and, you know, free labor at the same time. I think we're either gonna not create general superintelligence for survival

purposes and then we'll have useful tools to make people more productive, more creative.

[00:30:40] But I doubt we'll get friendly superintelligence with methods we're currently using.

[00:30:46] **Nate Hagens:** Is there a chance that some of the owners, the CEOs or the shareholders or the, the programmers at the highest levels of OpenAI and Anthropic, and elsewhere are building in controls of how to control the models that normal people wouldn't be aware of, including you, or is it just impossible to do that?

[00:31:13] **Roman Yampolskiy:** It would be wonderful if they figured out how to control them. That would solve the whole problem. I don't think they know how to do that. That's the issue. Yeah. So it doesn't matter who builds it. Could be the US, could be China, could be any company. It's the same uncontrolled outcome at the end.

[00:31:28] **Nate Hagens:** Well, it's, it's different dystopias then because if they were able to control it, then there's a couple humans that control everything, which might be slightly better than 99% chance of extinction, but still not a great outcome probably.

[00:31:44] **Roman Yampolskiy:** it depends. Once they do that, let's say it's hypothetically possible, they don't really need you for free labor or anything. They got AI, so it could be quite nice for you.

[00:31:53] **Nate Hagens:** But why would they care about me at all?

[00:31:55] **Roman Yampolskiy:** You need an audience, you need someone to follow your posts on Twitter. You need people.

[00:31:59] Somebody has to be impressed with your wealth.

[00:32:02] **Nate Hagens:** I'm, yeah, I'm a little more skeptical on that. But that's a hopeful thought. So there is- I'm an

[00:32:09] **Roman Yampolskiy:** optimist, you know.

[00:32:11] **Nate Hagens:** Yeah. Yeah. This is so profound. It, takes me a while, I'm a natural scientist, not a computer scientist, to feel the things you're telling me in my body, and I'm worried about inequality and resource shortages and everything, and the quality of the argument of what you're saying is just something that takes-- Well, it took you 12 years to process it or something.

[00:32:39] It's-

[00:32:40] **Roman Yampolskiy:** Still processing. We're still learning. Yeah. Still not, finished.

[00:32:44] **Nate Hagens:** Do you have children?

[00:32:45] **Roman Yampolskiy:** I do.

[00:32:46] **Nate Hagens:** Yeah. And are they aware of these existential risks?

[00:32:49] **Roman Yampolskiy:** Absolutely.

[00:32:50] **Nate Hagens:** So there is a related concept that I've been, exploring myself, something I term Goldilocks technology, not too hot, not too cold, that hits the sweet spot of midwifing us towards better civilizational trajectories, while also being appropriately matched to energy, ecology, social conditions that we're likely to face.

[00:33:14] Is there a technological sweet spot that can be found within artificial intelligence? I think you've alluded to it, but, like, what, would that look like?

[00:33:25] **Roman Yampolskiy:** Yeah, so we stop all training of general super intelligence, meaning we don't train on all the data. We don't try to make it as capable as possible, as soon as possible.

[00:33:34] We pick specific problems. Let's say protein folding is a great example, and we train an advanced model to do that one job. I'm not a philosopher. It doesn't drive cars. It doesn't play chess. It folds proteins. That's all it does. There is a human who will use that tool to be more productive, more creative, solve problems, cure diseases.

[00:33:56] There's lots of problems, no shortage of diseases. Pick one, monetize it, be rich and happy

[00:34:06] **Nate Hagens:** So I'm very naive on this topic, but it sounds to me, given the players involved, that this would have to be some sort of an international cooperation, like banning CFCs back in the day or something like that, because one country i– isn't going to unilaterally do this while giving up potential power, or if the other country gets to AGI first, they turn off, all your nuclear and financial things.

[00:34:34] So it seems like the only path forward for the type of slowdown and or even stoppage that you're suggesting would have to require the United States and China at a single table knowing deeply the things you're saying and constructing a path forward that is safer.

[00:34:56] **Roman Yampolskiy:** That's an ideal scenario, but I think the argument, if they want to keep power, then they should not build superintelligence.

[00:35:03] That's the realization.

[00:35:04] **Nate Hagens:** Because it's a dead end and

[00:35:07] **Roman Yampolskiy:** no one's- 'Cause they're gonna lose power, whether it is Communist Party of China or Tr- President Trump in the US, if you create something more capable than all of humans combined, you're not in charge anymore.

[00:35:17] **Nate Hagens:** And you think that something could arrive in twenty twenty-seven?

[00:35:22] **Roman Yampolskiy:** I think the recursive self-improvement process can start around that year. I don't know how long it will take to fully blow up.

[00:35:30] **Nate Hagens:** So here's my, from your perspective, maybe optimistic angle. AI itself requires a growing amount of energy, metals, water, and other materials which are all finite, which is the center of my work and has been for twenty years.

[00:35:48] Despite these physical constraints Most projections from Wall Street and the cheerleaders of AI and super intelligence assume that computers just simply keep scaling. Do you think there's a chance that the concerns you're presenting here could be prevented and preempted by resource and financial constraints, and that we, at least for a while, go into an AI winter, and that because of resource constraint reasons, buys people like you time to, to slow down the development?

[00:36:24] **Roman Yampolskiy:** So obviously, there are fixed resources. We don't have an infinite supply of energy or anything else, but we are so close to human level, we will definitely not hit that limit before we get to human level and above.

So will it become a larger part of our economy? Yes. Are they also becoming more efficient?

[00:36:48] Are the costs of tokens basically collapsing? Yes. So I don't think it's going to slow it down enough or in time to give us decades of thinking time.

[00:36:59] **Nate Hagens:** How do you integrate the high cost of some of the United States models and the incredible circuitous financial, creative ways of Oracle and Nvidia, doing circular finance models and such like that relative to ten percent the cost of some of the, the Chinese models.

[00:37:27] is that relevant to this story or are both of them just headed over the cliff?

[00:37:32] **Roman Yampolskiy:** So there are different ways of doing it. There are dangers of open source models. You're giving intelligence weapons to psychopaths. That's not optimal, but more efficient, I guess. I think they achieve some of this efficiency because the Chinese government is sponsoring some of that work and helping them, so it may not be fully sustainable.

[00:37:52] at the same time, well, yeah, there is circular finance, but also Nvidia generates a hundred billion dollars. It's not a fake company, right? There is lots of need for this, not just in AI, but, obviously all sorts of digital services, cryptocurrencies, everything needs to be computed.

[00:38:09] **Nate Hagens:** So I understand from our mutual friend that you came into this field to research whether we could build AI that was safe and beneficial for humanity.

[00:38:20] but as you've described, you have discovered there are even more open questions about the nature of intelligence yourself-- itself in addition to the

risk that you've outlined. And you have previously proposed a novel term, intellectology, to refer to the study of the forms and limits of intelligence. And I wonder, Roman, at what point in your research journey did you begin to see the need to distinguish intelligence from wisdom, and what led you to make that distinction?

[00:38:56] **Roman Yampolskiy:** So unlike AI safety, intellectology is not a popular term. Nobody picked it up. I'm still hoping. But the idea is that we have many fields working in essentially the same part of a problem, but using different tools, different vocabulary. Often we don't know about each other. So you have artificial intelligence studying intelligence in certain substrates, but then you have neuroscience, you have psychology, you have people studying consciousness.

[00:39:22] But I think all of it is just different subdomains in the study of intelligence. What can be intelligent? How intelligent? Different types of intelligence. Can they be conscious? How do we measure intelligence, detect it? How do we tell if an output is produced by an intelligent process or natural process?

[00:39:40] All of it is intellectology, and I think it's the most interesting area of research, and everyone's kind of working on it just in isolated subdomains.

[00:39:51] **Nate Hagens:** It sounds a little like E.O. Wilson's concept of consilience. I

[00:39:56] **Roman Yampolskiy:** have to look it up.

[00:39:57] **Nate Hagens:** That was my Bible like 25 years ago. So in some ways, Homo sapiens, wise man, we have not been, but we have been intelligent and clever, and this AI tool is just a manifestation of the left brain, restless, dopamine conquest, discover, novel part of our phenotype.

[00:40:27] And I think the pathway forward, if there is going to be a long run for humanity and the biosphere, lies more in wisdom. Can AI m-models, either the ones that I use or ones in the future offer wisdom in that sense, or are they just downstream of the weightings and the things that created them, which are intelligence based?

[00:40:58] I don't know if that makes sense.

[00:41:00] **Roman Yampolskiy:** I think if you can formalize what you mean by wisdom, you can propose how to measure it, then AI can definitely excel at it and beat humans at it.

[00:41:08] **Nate Hagens:** Excelling at wisdom Yeah. Okay. I'm gonna, I'm gonna think about that one. So what are some sources you draw wisdom from, in, in your own life, Roman, that help you understand our present circumstances and the future?

[00:41:25] **Roman Yampolskiy:** I recently started a podcast. I follow some great minds in that Roman Forum. Here we are.

[00:41:32] **Nate Hagens:** Okay.

[00:41:32] **Roman Yampolskiy:** And, I try to talk to the world's smartest people, slowly growing my number of episodes to more and more of them. I think human intelligence is still a great source of wisdom, and I'm trying to get to the best.

[00:41:45] **Nate Hagens:** Human kindness, empathy, sense of humor, unexpected, shift in the discussion to a different topic, all those things, Suggest to me wisdom, intelligence or a focus on-- So last week's episode on my podcast was with a mathematician named Greg Elliott, who just wrote a book called The, The Psychopathic Selection Hypothesis, and he didn't talk about human individual

psychopaths, but that our entire economic system is now tilted from rules, laws created in the past to act in an instrumental way, where the number that we're optimizing is become more important than the thing it was meant to represent.

[00:42:43] and that to me seems like intelligence optimized for the wrong thing, and now we have eight billion humans that are trying to be good people, and pro-social, and kind, and generous, and selfless, and respectful of the biosphere. But we're sitting in this psychopathic economic system, and to me it seems like we're bolting on a new software upgrade to this system with AI and the things that you're suggesting.

[00:43:19] Do you have thoughts on that?

[00:43:20] **Roman Yampolskiy:** There are definitely some problems I notice with the system, so I get to interact with a lot of very, very successful individuals in terms of money accumulation. And interestingly, none of them can tell me what they do with money after a certain amount. So with the exception of Elon Musk, who has a specific plan to build a city on Mars and needs a trillion dollars for that, I get that.

[00:43:44] But everyone else who has billions of dollars, I don't think they have any clue what an additional billion does for them. It's more like an addiction to collecting zeros in your account.

[00:43:54] **Nate Hagens:** Well, it's an addiction to power and aversion to shortfall risk because our society optimizes for that, and money is the ultimate optionality because you can turn it into anything else, including ownership in a tier one AI play.

[00:44:10] But I think it will end badly, in the same way that you said, no matter what we do with AI, if it ends and you're losing all your power, you lose all your

power. yeah. No, that's interesting. So you earlier suggested AI combined with robotics might eventually automate a significant, if not overwhelming majority of jobs.

[00:44:34] On this show, we talk a lot about what humans are for beyond their roles as workers and consumers. And if work stops anchoring people's identity, where do you think meaning might come from and what should societies be doing now to prepare for the few pathways that don't result in extinction and do result in a lot of robots and not a lot of humans working?

[00:45:03] **Roman Yampolskiy:** Again, setting aside the existential risk, suffering risk, now we talk about I risk, Ikigai risk, a risk of losing meaning. We can look at a population of people, we call it retirees. They no longer have to work. They have some unconditional basic income. What do they do with their time? You can go socialize.

[00:45:22] You can go fishing. I think virtual worlds will offer a lot of opportunities to do whatever you want in novel domains. If you have a substrate of intelligence controlling your virtual environment, you can have a lot of fun exploring universes, meeting aliens. So it really depends on what you're into. I don't think we're going to be bored.

[00:45:40] I'm always puzzled by people who say, "I don't want to live longer because I'll be bored." To me, it's moronic.

[00:45:48] **Nate Hagens:** So beyond AI itself, that we are opening the proverbial Pandora's box, continues to be integrated further and further into our daily lives, just even from six months ago. How do you think your research could or should change the way we develop all technology, not just AI, and think about technology's role in the future we're trying to create?

[00:46:18] **Roman Yampolskiy:** So the general rule is don't create something you don't control. Don't create something capable of wiping out large numbers of humans. We talk about gain of function in AI, but it's the same problem with gain of function in biology and viruses. Don't do that type of experiment.

[00:46:33] **Nate Hagens:** So it's almost like humanity has to start operationalizing the precautionary principle when they invent or aspire to something.

[00:46:46] We've not been so good at that.

[00:46:48] **Roman Yampolskiy:** No, we haven't.

[00:46:50] **Nate Hagens:** Yeah. So I have some closing questions that I ask all of my guests, but I'm not sure that I've completely plumbed the depths of your own wisdom and expertise on these things. What-- can you summarize, what you would like the average person listening to to take away from your 15 years and counting of deep research on this topic?

[00:47:26] **Roman Yampolskiy:** Don't build a replacement for humanity. Don't make yourself obsolete. Develop useful tools to make your life better, everyone's lives better, more creative. We can have abundance, we can have longer health spans, lots of opportunities with technology, but we have to create tools, not agents to replace humans.

[00:47:45] If you personally don't have any power in that space, maybe you know someone who does. Maybe you can vote for someone who can influence this direction. Do what you can.

[00:47:55] **Nate Hagens:** So we have an election coming up in just two months. From your perspective, you would say, I'm guessing, that this is one of the central issues of the next year or two?

[00:48:08] **Roman Yampolskiy:** It should be the only issue every political party is discussing. It's not even on the radar for most of them. It's insane.

[00:48:15] **Nate Hagens:** Do you have your equivalents, I imagine in China, where there are people cautioning the Chinese government about the path, or is it a different sort of setup there?

[00:48:26] **Roman Yampolskiy:** I understand there are Chinese academics who are actually meeting with American academics and international counterparts to kind of figure out what to do, and if Chinese counterparts do it, that means Communist Party authorize those negotiations, so I think it may have some opportunities, to result in productive outputs.

[00:48:46] **Nate Hagens:** If you had to guess, How would it become possible for six or eight very senior, US government people to meet, and some of the, Anthropic or OpenAI people to meet with their counterparts in China to, come up with a plan to take what you and your colleagues are saying seriously? How would we accelerate the chances of that happening?

[00:49:16] **Roman Yampolskiy:** I think it's already happening. There is something called dialogues between Chinese Academy of Science and US counterparts, and Canadian counterparts. I don't know the latest state of the art, but they produce periodic reports. You can see what was discussed and what they agreed on.

[00:49:30] **Nate Hagens:** Yeah. Okay. so if you have a few more minutes, I– Sure

[00:49:36] I have some personal questions that I ask all, my guests. Do you have any personal advice to the people listening at this time, where they're aware not only of climate change and polarization and AI, the risk you bring up today, what some would call the, the meta crisis? Do you have any personal advice for being alive at this time?

[00:49:58] **Roman Yampolskiy:** So it doesn't matter how much you have left, it's about collecting experiences, and if you collect it today, it's as valid as if you have fifty years or a hundred years. So enjoy life, collect the best experiences you can.

[00:50:10] **Nate Hagens:** And you mentioned you had children and you are a college professor. Are you a researcher or do you actually teach students?

[00:50:17] **Roman Yampolskiy:** I actually teach students. I teach AI.

[00:50:20] **Nate Hagens:** You to undergrads?

[00:50:23] **Roman Yampolskiy:** everything, graduate, PhD, all levels.

[00:50:26] **Nate Hagens:** I taught at the University of Minnesota a class called Reality 101 for nine years, and I really miss it.

[00:50:32] **Roman Yampolskiy:** Cool title.

[00:50:33] **Nate Hagens:** Yeah, Reality 101: A Survey of the Human Predicament. And I used E.O. Wilson's, Social Conquest of Earth, as the main textbook.

[00:50:41] **Roman Yampolskiy:** What department offered that course?

[00:50:44] **Nate Hagens:** That's a very astute question. No department could have offered it, so it was in the honors college, which was a general, you know, honors elective course.

[00:50:55] **Roman Yampolskiy:** Interdisciplinary.

[00:50:56] **Nate Hagens:** Exactly. Yeah. It was interdisciplinary, yeah. Because it wouldn't have been approved in the economics department, as one example.

[00:51:04] So, what recommendations do you have for young humans in their teens and twenties who become aware of all this stuff?

[00:51:10] **Roman Yampolskiy:** That is super hard. I have a seventeen-year-old basically going to college next year, and I have no idea what makes sense ten years from now when he gets his PhD in whatever. So if there is something you just love learning about, it's one thing, but if you're doing it strictly to get a job, I would consider starting a company instead.

[00:51:30] **Nate Hagens:** What do you care most about in the world?

[00:51:33] **Roman Yampolskiy:** I want to know what's true, what's real. So if it means hacking the simulation to get to real knowledge, so be it.

[00:51:40] **Nate Hagens:** If you could wave a magic wand, and there was no personal recourse to your decision, what is one thing you would do to improve the future for humanity and the biosphere?

[00:51:51] **Roman Yampolskiy:** I think you know my answer here.

[00:51:55] **Nate Hagens:** You would stop AI cold, on the development towards superintelligence right now.

[00:52:02] **Roman Yampolskiy:** Superintelligence, right. So the AI term, I like technology, I like science, I like engineering. Yeah, Develop your tools, God bless you, but don't build gods.

[00:52:10] **Nate Hagens:** Well, this has been an unusual conversation for my podcast.

[00:52:14] I'm left with the feeling that the antidote to some of the things we face, including AI, but not limited to AI, is humans like you. because you're no BS, and

you're just honest. You lay it out, and you care, and there's something palpable about that. So I appreciate your time today and your work.

[00:52:40] **Roman Yampolskiy:** Thank you so much. I appreciate you having me.

[00:52:43] **Nate Hagens:** Spasiba. We'll talk soon. If you'd like to learn more about this episode, please visit thegreatsimplification.com for references and show notes. From there, you can also join our Hilo community and subscribe to our Substack newsletter. This show is hosted by me, Nate Hagens, edited by No Troublemakers Media, and produced by Misty Stinnett and Lizzy Sirianni.

[00:53:08] Our production team also includes Leslie Batt-Lutz, Brady Heyen, Julia Maxwell, Gabriella Sleiman, and Grace Brunfelt. Thank you for listening, and we'll see you on the next episode