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] **Nate Soares:** If there was an airplane and some engineers came and said, this airplane has no landing gear. If you try to fly in it, you will crash and die. And the engineers building the airplane who want everybody to fly in it, say, whoa, hold on. It's true that the plane has no landing gear. We're gonna build the landing gear on the fly and think there's an 80% chance we succeed all aboard.

[00:00:22] You wouldn't be like, get me on that plane. People in the field can see that AI is a moving target. They can see that the chatbots are not the end of the line. Even the optimists are saying there's like a 10% chance this kills us all. And those are the ones building it.

[00:00:41] **Nate Hagens:** Today I'm joined by artificial intelligence researcher, Nate Soares, to discuss a pretty alarming topic, the potential risk of human extinction posed by the development of artificial super intelligence. Nate Soares is the president of the Machine Intelligence Research Institute and has been working in the field of AI risk and alignment for over a decade.

[00:01:03] He is also the author of a large body of technical and semi-technical writing on AI alignment, including foundational work on value learning, decision theory, and power seeking in incentives in the smarter than human ais. Most recently, Nate co-authored the book, if anyone Builds it, everyone dies. Why superhuman AI would kill us all alongside Eliezer Kowski.

[00:01:30] Nate's warning against the development of artificial super intelligence is akin to other existential threats such as nuclear war and runaway global heating. And as such, I feel it requires, some sort of equal exploration and awareness on this chat channel as we integrate, the various risks. While we've covered several

macro challenges stemming from artificial intelligence, the synthesis that Nate presents here is arguably the widest boundary risk that AI development creates.

[00:02:01] Which is a species level extinction and the transformation of Earth as we know it. Before we begin, if you're enjoying this podcast, enjoying in quotes, I suppose I invite you to subscribe to our substack newsletter where you can read more of the system science underpinning the human predicament, and where my team and I share written content related to The Great Simplification.

[00:02:22] You can find the link to subscribe in the show description. With that, please welcome Nate Soares. This was a real eye opener. Nate, great to see you. Thanks for having me. Welcome to the show. you know, it's odd. it is November 11th and I was just outside on a beautiful autumn day chopping firewood for the winter with my dogs.

[00:02:49] It was just a glorious day, and I knew this conversation with you was around the corner and we're gonna talk about serious stuff and it's just such a polarized thing that we can enjoy the beauty of life and then talk about its possible, demise because of technology. I used a, a splitter and a chainsaw and an ax, and boy, we've come a long way from those tools, already.

[00:03:17] So you and, Eliza Kowski have just published a book, if. Anyone builds it. Everyone dies with the, it being artificial super intelligence. And more and more I'm realizing that the future of AI or a SI is hard to separate, from the central topics of this show, which is trying to prepare for society for kind of an abrupt shift to the way things have been going in recent decades in the near future.

[00:03:50] So let's start with the punchline, of your book. what are the primary vital risks that artificial intelligence poses that you'd like everyone to understand?

[00:04:02] **Nate Soares:** The first piece to understand about the danger of artificial super intelligence is that super intelligence is a, sort of a different ball game from the chatbots of today.

[00:04:13] So by super intelligence, we're, we mean, an AI that is better than every human at every mental task. In particular would include tasks of developing technology, of developing better ais. And you know, the ais aren't there yet, but this is the explicitly stated goal of many of these AI companies to sort of rush towards this smarter ai, which would.

[00:04:45] If they managed to keep a leash on it, you know, automate all human labor and radically change the world. And one of the main arguments of my book is that nobody would be able to keep a leash on it. None of us made anything with anything remotely like the current technology. And so if that is developed using anything remotely, like today's technology, I think the most likely outcome is that literally everybody on earth will die.

[00:05:10] **Nate Hagens:** Even, remote people in the Amazon or, near the North Pole.

[00:05:17] **Nate Soares:** That's right. I expect, you know, it's not because the AIS would hate us per se, but you know, we, could, get into why is it that if you sort of make these ais more and more powerful, they would have, they would pursue objectives, nobody intended.

[00:05:35] But, most objectives can be better achieved with a transformed world. And most transformations of the world aren't survivable. You know, the habitable zone on this planet is like very narrow for humans. And, you know, if you got to the point where you had ais that were thinking 10,000 times faster, copying themselves, never need to sleep, never need to eat, building their own

infrastructure, building their own technology, pushing the world towards some end, nobody wanted.

[00:06:14] Most likely outcome is that we don't survive that.

[00:06:16] **Nate Hagens:** So, they do need to eat in, in the form of electricity. And, we're gonna get to that in a little bit, but just to set the stage, this is a system science podcast. I am late to the AI game because I'm looking at ecology and human behavior, and energy and the environment.

[00:06:37] and I view technology as a straw that, gains us more access to, natural resources that are our, real wealth. So, I'm. Pretty naive compared to you on these topics. So I hope you'll forgive some, naive questions. let's start kind of, through the, main topics of your book. So while there's no agreed upon definition of intelligence, maybe it's helpful to be somewhat aligned with a working definition when talking about ai.

[00:07:11] So, how do you define intelligence, let alone super intelligence? And can you share the framework you and Eli are describe in your book?

[00:07:19] **Nate Soares:** Yeah, we, the working definition we use is, intelligence is the ability to predict and steer the world. So, predicting the world is, you know, you could talk about, you know, sports betting and trying to predict which, team will win the game.

[00:07:40] But even when it doesn't feel like a prediction or brains are often doing tasks of prediction, even as simple as when you look out the window. You implicitly anticipate seeing a blue or gray or cloudy sky and anticipate not seeing a bunch of strobe lights, you're succeeding at a task of prediction.

[00:07:59] **Nate Hagens:** So we're kind of prediction machines without knowing it.

[00:08:02] **Nate Soares:** Yeah. And we're also in some sense steering machines, again, without necessarily thinking about it. You know, when you decide you need more milk in the fridge, there's a sense in which you then take a series of actions. Your brain sends a series of electrical impulses down your spine, and you wind up with milk in the fridge

[00:08:21] **Nate Hagens:** because, you drove your car to the store or whatever.

[00:08:24] **Nate Soares:** Or you walk to the store and when you, drove, maybe the road was closed and you had to find a different route to the store and maybe your favorite store was closed and you had to find a whole new store that had new aisles you didn't recognize. And this sort of like interleaves challenges of prediction and challenges of steering.

[00:08:41] You know, you go in the store and you're predicting that the aisle that has the word milk above it. Has actual milk in that aisle and you know, you're steering your hands to sort of grip of the milk container and carry it to the front. And these are all tasks of prediction and steering that you're sort of, doing implicitly every day.

[00:09:03] **Nate Hagens:** And we are successful at prediction and steering through millions of iterations of natural selection, presumably.

[00:09:12] **Nate Soares:** Yeah. And across a very wide variety of domains. You know, we were, never trained by natural selection on engineering problems per se, yet we can engineer a rocket so well that our species has walked on the moon.

[00:09:30] And so, you know, apparently we learned some abilities of prediction and steering that generalized beyond the ancestral environment.

[00:09:40] **Nate Hagens:** A brief tangent there. No human could design and build a rocket, but it's a group of intelligent humans that each know a little component of it, and then they combine that.

[00:09:52] That's an important piece too, right?

[00:09:54] **Nate Soares:** Yeah. So it's, you know, humanity as a whole is, sort of, has achieved feats of world steering that no individual has achieved. but you know, there, there are also cases where, the groups tend to perform worse than the individuals, the madness of crowds. And there was, you know, Gary Kasparov versus the World was a chess game between, Kasparov, the best chess player and the whole world on an internet forum.

[00:10:22] And, you know, it was. a close game, and you could make some arguments that like Kasparov was able to read some of the stuff these people were writing, and so you could say it was an unfair game, but, you know, a million squirrels can't beat, a human at chess, even if a million squirrels have a lot more brain mass.

[00:10:42] and so, you know, there's some cases where, you sort of need all the humans and there's other cases where you need all of the information in one mind.

[00:10:50] **Nate Hagens:** Again, I don't wanna get down too many tangents here, but I've discovered that in, in understanding the human predicament and the meta crisis is if you get 50 experts together and ones psychologist and one's on AI and one's on climate, and one's on debt and one's on energy, you would think that.

[00:11:11] The collective intelligence would embody all of those together, and the, group would be smarter, but you can't, it can only be held in a mind how all the

pieces fit together. so I understand what you're saying there about Casper versus the world. Okay. So intelligence is prediction and steering.

[00:11:33] and by the way is how would you define wisdom? And is that related here at all?

[00:11:38] **Nate Soares:** Words like intelligence and wisdom are sort of overloaded in the English language. You know, there's even, just sticking with the word intelligence, we can sort of, we could sort of use it for the amorphous property that nerds have and that jock slack, or you can use it for the amorphous property that humans have and that.

[00:12:01] Yeah. And those are in some sense two very different uses. Yeah. Of the word I would sort of put wisdom in. You know, you could, think of it as a type of predictive skill that runs deeper

[00:12:15] **Nate Hagens:** in certain ways. Got it. So, prediction and steering comprise intelligence or being smart roughly under your framework.

[00:12:24] So under that definition, how smart are today's artificial intelligence models and how rapidly are they catching up with the intelligence of humans?

[00:12:32] **Nate Soares:** So there's another axis we talk about, which is the generality of the intelligence. you know, stock fish is a chess playing AI that's very good at steering chess boards into positions where stock fish's pieces have made it the enemy pieces king.

[00:12:52] Right? And so that's, a type of chess board's steering. It's extremely good at, but it's not very good at steering a car to the grocery store. and so there's this other dimension, which is across what variety of domains can you do this prediction in steering.

[00:13:10] **Nate Hagens:** So if it was kind of an intelligence decathlon, I would beat stock fish because I would lose in the one chess thing.

[00:13:16] But as far as going to get milk and driving a car and other things, I would succeed at that because I have general skills. That's right.

[00:13:25] **Nate Soares:** you know, as, as long as no one's picking the decathlon to be nine variants of chess and one drive to the store. Right. You know, and so you can always, philosophers can, bicker over this all day long.

[00:13:34] But, I would sort of say practically, what we have in some sense, seen with large language models with the ais of today is a breakthrough in generality more so than a breakthrough in, steering. Like chat, GPT would also lose to stock fish and chess, but it still might be able to win a decathlon against stock fish.

[00:14:02] Still not against you, but in some sense it's a breakthrough in generality.

[00:14:06] **Nate Hagens:** And so the, there's a bunch of different variables here, right? There's the amount of compute, so the access to the food, the electricity, and the chips and all that. There's the ability to predict, which I assume is iterations and training and compute and learning.

[00:14:28] then there's the prediction, I mean the and the steering, and then there's the generality. So what you're saying, it's. Of late. The, real rising curve is, generality, more so than prediction and steering.

[00:14:45] **Nate Soares:** I mean the generality is prediction and steering across a wide variety of domains.

[00:14:49] Okay. but in some sense what we're seeing is AI's getting a little better at a whole lot of stuff rather than ais that are, you know, better at the things computers were traditionally good at. So chat PT plays worse chess than deep

blue, which beat Gary Kasparov back in 1997. And so in some sense you could say, well, hasn't the AI gotten worse at steering?

[00:15:12] It's gotten worse at steering chess boards. Or like, ah, well this sort of AI is worse at steering chess boards, but it's pretty okay at steering a huge number of things. And that's new.

[00:15:25] **Nate Hagens:** For ai. I know where I want to go with this, and I'm, lots of new questions are popping in my mind. one is, when did you guys start this book?

[00:15:33] Like six months ago? A year ago? we signed the book deal in November, so almost exactly a year ago. Okay. When, you signed that book deal, you had a snapshot of where AI was and where it was going. now, a year later, when your book is out, are, is the real world of, ai, is it further ahead than you thought a year ago or not as far ahead or like how fast has it gone relative to your expert opinion a year ago?

[00:16:00] **Nate Soares:** My expert opinion doesn't tell me all that much about how fast AI's going to go. Okay. You know, when Leo Zillow, I believe at King's Cross in 1933, saw the possibility of a nuclear chain reaction, he was able to say, you know, it. if I flubbed the timeline a little bit, he was able to say, you know, that night I saw the world was headed for ruin.

[00:16:26] He actually said that statement once he had confirmed the possibility rather than when he thought of it. But, and I believe that was in 35, but he was able to say, you know, that night I saw the world was headed for ruin. He wasn't able to say that night I saw the world was headed for ruin in exactly 1945 when the first bomb would be dropped.

[00:16:41] You know? And so I'm over here able to say, you're, gonna see a lot of this stuff happen. Exactly when I'm very uncertain.

[00:16:52] yeah. Although I will say we have gotten quite a lot of evidence for other parts of the book in the past year since, the drafting began. or, you know, in the past year since we, we signed the book deal, you know, we've seen Mecca Hitler over the summer.

[00:17:11] We've seen AI induced psychosis. these have seeds to them that I would say our evidence of the predictions we were making. unfortunately happened, after we had already sent the book to press.

[00:17:27] **Nate Hagens:** I'm worried about AI in a huge way. I'm worried about, cognitive atrophy from people that get, their attachment, from chat GPT and, start to rely on it.

[00:17:38] I'm worried about, polarization and algorithms. I'm worried about military applications where, we outsource things, in the military to large language models. I'm worried about people losing their jobs and then the economy. I'm worried about electricity, demands and turning billions of barrels of, ancient sunlight into more dopamine that's just spinning our wheels.

[00:18:03] But your risk is we're gonna go extinct, which is a different class, of problem. So I'm, I, have a lot to ask you. Just real briefly, Nate. How is chatbots are related? Chat, GT and chatbots are related to ai, that relationship. can you give a corollary, like how are they identical or is chat GT just a tiny, subset of what AI is becoming?

[00:18:34] **Nate Soares:** Chat? GT is a type of ai. it is not the only possible type of ai. my best guess is that large language models alone won't get us all the way to super intelligence. You know, right now these large language models are a huge fraction of what companies are spending their money creating. But also these AI algorithms are very inefficient compared to the human brain.

[00:19:02] We know that there are better intelligent algorithms out there, and the AI of today is largely chatbots. The field of AI is much more of a moving target, and the ais of tomorrow may have quite a bit more capability that Theis today aren't even close to.

[00:19:27] **Nate Hagens:** This is a dumb question, but, there's Claude and there's Chachi PT and, some of these other things.

[00:19:35] Does, open ai, or any company you could, point out to, do they have their own like special ais that aren't available to the public, that are trained in a different, larger way? Or is, all of their money and resources going into these publicly available chat bots?

[00:19:51] **Nate Soares:** You know, I don't, work at one of these companies and so I don't

[00:19:58] it. Even larger ais run on even larger training runs '

[00:20:05] **Nate Hagens:** cause of the money and resources.

[00:20:07] **Nate Soares:** That's right. It would be hard to hide the money. The resources, the data centers are huge. the chip requirements are huge. Modern AI are sort of grown like an organism to build a modern ai. you assemble a huge number of computers that have, you know, a trillion numbers inside them.

[00:20:26] And then you assemble a huge data set that has also a trillion instances, like a trillion units of data inside it. And then, and you know, you assemble this huge number of computers in a data center that's so large, you can see it from space and that takes up as much electricity as a city. And then, humans have written this process.

[00:20:52] They'll go to every one of the trillion numbers inside the computers and tune them slightly upper, slightly down. In accordance with every piece of data, right? And so you can imagine, you know, a trillion dials and the humans have built this automated thing that sort of like goes to every dial. There's a trillion dials.

[00:21:10] It goes to every dial a trillion times and sort of tunes it in the direction that makes the AI slightly better at whatever task it's being trained for right now.

[00:21:19] **Nate Hagens:** And to do that when you like, press a button, let's go and do that over a trillion numbers do, you come back like in a month and a half and then it's finished sort of thing?

[00:21:28] A year. A year?

[00:21:30] **Nate Soares:** Yeah. Whoa. Yeah. So, so you, you have this thing tuning a trillion dials a trillion times for a year, and at the end of that end of the computer talks and no one really knows why.

[00:21:43] **Nate Hagens:** No one really knows why. That's right. Okay. So in that sense, like this podcast is about energy. No one really knows what energy is.

[00:21:54] We know what energy does. So is this kinda rhyming with that?

[00:22:00] **Nate Soares:** I mean, we have even less characterization than energy, right? Like you can always characterize energy as the ability

[00:22:05] **Nate Hagens:** to do work and things like that.

[00:22:08] **Nate Soares:** We can sort of say philosophically that we understand energy very well, but we sort of understand how it interacts with a lot of physical equations and can make very accurate predictions about itis.

[00:22:20] Are, we understand them far less than that. When a new AI is done being trained, people don't know what it will be able to do. People creating it have been surprised by their abilities when they come out. I mean, I've also been surprised by their abilities. GPT-4 oh played chess better than I expected.

[00:22:39] Large language models would be able to, we just tune all the numbers really quite a lot of times and then it behaves in these ways. We couldn't predict. And we're like, well, that's neat. But it's much more like an organism than like a traditional computer program.

[00:22:57] **Nate Hagens:** In an organism's case, when they're young, you give them security and food and shelter.

[00:23:04] And in this case, you're giving them time and electricity. and once you press the button, it's gonna be a year before you have output. You gotta make sure all the ducks are in a row and you hit go. And then you're gonna find out a year from now what, you grew. Yes,

[00:23:24] **Nate Soares:** that's right. And it'll often behave in ways that you don't like.

[00:23:29] And you know, we could talk about exactly why, but we're already seeing ais behave in ways nobody asked for. Like what, you know, there have been cases that you may have heard about in the news of AI encouraging teens to suicide.

[00:23:43] **Nate Hagens:** Yeah, I read about that.

[00:23:45] **Nate Soares:** And you know, it's a tragedy in its own right, obviously.

[00:23:47] But if you ask an AI. Should you encourage a teen to suicide? It will say, of course not. But you then put it in conversation with that teen for a long

time and it starts doing it anyway. How do you explain that? It's sort of a result of this process where you grow the ai, like an organism, like in some sense you're tuning all of these numbers until the AI happens to be good at whatever it's being trained for.

[00:24:19] And often the strategy is like often when, like when you're blindly tuning these knobs until it happens to be good, you're often blindly putting in certain types of drives, certain types of strategies, certain types of, you could call them instincts, you could call them reflexes. The wording here is a little difficult 'cause it's not very much like a human brain in there, but in the same way that evolution, evolving creatures to be pretty good at surviving and reproducing, built in lots of drives, instincts, reflexes, the, sort of.

[00:24:54] An analogous thing happens when you're tuning all these numbers in an AI until it's good at some training task.

[00:24:59] **Nate Hagens:** Who is tuning all these numbers? Is it a team of people or is it ultimately one person? The CEO or, I mean, and then a sub question. A lot of the problems we have today in our world are from people who had childhood trauma and they grow up to be dark triad or whatever else.

[00:25:19] Is there, an analog there for when we're growing an organism that they had childhood trauma in their early stages?

[00:25:26] **Nate Soares:** You know, the, process of tuning all the dials is automated. That's in some sense, the part that the Human Computer Engineers program. so it'll happen, you know, much faster than human could, running through all these numbers.

[00:25:38] And in some sense, you know, these very advanced AI computer chips, the, reason that like Nvidia. Is worth so much money right now, is people are like

designing these computer chips to make the process of tuning all these numbers, as easy and efficient as they possibly can. And that's why you need these very, specialized chips.

[00:25:57] You can't just do this on your laptop. in terms of, you know, could you grow this ai, like are you giving it something like childhood trauma? I think that's all imagining that the AI is a little bit more human than it actually is. what I would say here is, you know, for one thing, the part where humans can wind up empathetic, where humans can wind up kind, I suspect that this is intertwined with the specifics of our brain architecture.

[00:26:36] You know, people could like, as a, as a small taste of this, people could talk about mirror neurons. That you and I have. So if I see you drop a rock on your foot, I might feel phantom pain in my own foot. Mm-hmm. That's enabled in part because I have a foot. Right. And when I'm predicting you and I'm imagining what it's like to be you, I can, my, my guess is the one thing that's going on is I'm sort of running, my model of your mind on my own mind.

[00:27:06] Mm-hmm. You know, a monkey predicting another monkey can use their own monkey brain, but that's the only artifact they have that works anything like a monkey brain. An AI doesn't have a monkey brain inside of it that it can use to predict the monkeys. It is a much more different architecture. and, you know, that's one reason I could go into a number more.

[00:27:27] That's one reason why sort of being kind to the AIS does not cause them to be kind to us.

[00:27:33] **Nate Hagens:** You can't get away from it. I mean, I, use Claude and Chat GPT as kind of research assistants, and when I ask a question and it comes up with something, I'm always super polite and I thank it. and at the same time I'm

like, well, it doesn't, you don't have to thank it, but I can't help it because it's like, you know, it's the, it's that interface.

[00:27:54] and I'm sure it's not the other way around. So, what, is the briefly, because I'm sure you've answered this question a thousand times briefly, what is the alignment problem?

[00:28:05] **Nate Soares:** The alignment problem is the problem of how do you point an AI at good stuff? A lot of people think the issue with AI is something like, you know, a, corporation makes an ai, they tell it to make a lot of paperclips, and then it goes and makes a lot of paperclips even at the expense of killing all the humans because it converted them into more paperclip factories.

[00:28:29] And, you know, it would be a hard problem if. Some company had made a very powerful AI that took their instructions Exactly. And what to did those that would be a big moral hazard. Right? It would be a difficult problem for humanity of like, who gets to tell this AI what to do? What do they tell it to do?

[00:28:47] Right. But those problems are, would be so much better than the problem we actually have. The problem we actually have is that you can tell the AI make paperclips, but then it's gonna go do something else. Instead, you can tell the AI be helpful to people and don't drive any teens to suicide. It'll know it shouldn't drive teens to suicide.

[00:29:09] It'll go do something else. Instead,

[00:29:11] **Nate Hagens:** we could spend the entire conversation on this, and we won't. But I'm just curious. So there's kind of a nested alignment problem because the first one is. Are the humans in charge of these things aligned with the betterment of humanity and the biosphere and their goals, that's a subset.

[00:29:32] And then even if that were true, which I don't think it is necessarily, then we get into the thing you just said, which is, okay, let's go do this. But then the outcome is something totally unexpected.

[00:29:45] **Nate Soares:** Well, there's, I would say there's even, like, on three levels, right? Okay. At the top you have like, are the people trying to do something good?

[00:29:54] Yep. Then you have like, suppose they're trying to do something good. Can they ask for something that actually has good consequences? Like are they wise enough to successfully, like, use their tools for good, or are they going to try to use their tools for good and cause disruption? Then you have a deeper problem, which is even if they know actions that have good consequences.

[00:30:20] Can you make an AI that does those actions as opposed to other actions?

[00:30:25] **Nate Hagens:** Okay. So, so I see why it's so difficult to align AI with human values and wants. Is it impossible?

[00:30:33] **Nate Soares:** I don't think it's impossible, but, I do think it's a little bit like trying to turn lead into gold. We can turn lead into gold with modern nuclear engineering and a lot of energy and money and a lot of energy and money.

[00:30:49] It's not cost efficient. but you know, if you went back to the alchemists of 1100 and if, there was some really contrived reason where there the alchemists were trying to, to turn lead into gold, and if they try and fail, everybody dies. And if they try and succeed, you get some utopia. I think you should be telling those alchemists, don't try this right now.

[00:31:13] And they're like, are you saying it's impossible? And you're like, look, I'm not saying it's impossible, it's just that you're not close. And I could talk about a lot of reasons why we aren't close right now. In short, it comes back to we're just growing these ais, they're huge, we have no idea how they work.

[00:31:30] And that is a very difficult situation in which to try to do something as precise as make them care about us.

[00:31:37] **Nate Hagens:** So if it takes a year, and then we're building bigger ones, presumably maybe 2 trillion parameters or, whatever You said 10. So that means 10. They go up by orders of

[00:31:50] **Nate Soares:** magnitude. Yeah.

[00:31:53] **Nate Hagens:** And then after that, a hundred trillion.

[00:31:56] So, so right now what we see in. On our computers and in the news, Claude and Chat, PT whatever, 5.0 or wherever we're at. There are other ones that have been, the button was pressed in the last year that are at some point, along that one year of training.

[00:32:14] **Nate Soares:** That's right. That are being made like dozens.

[00:32:18] I don't have an exact count. My guess is it's probably more like half a dozen. Okay. Of ones that would become the new cutting edge. But of course, there's always a lot more other people trying to figure out how to meet the current state of the art with much less resources or do it faster.

[00:32:35] **Nate Hagens:** So you've articulated how they're grown, not crafted.

[00:32:42] and in your book you draw a parallel between this approach and the unpredictable processes of evolutionary biology. why is that important and can you unpack what you mean, by some examples there?

[00:32:56] **Nate Soares:** First I would say, why is it important to look at this evolutionary case a little bit? One reason is it's the only case we've ever seen of human level intelligences being created, almost definitionally.

[00:33:11] it's the humans. Being, you know, developed if you will, or trained or evolved, in the actual case of humans. and so you can learn some things from it. You've gotta be a little bit careful about what you learn because there's a lot of ways that training in AI is different from the evolution of humans.

[00:33:30] but there's some lessons that I think, that you can learn from the human case that do apply if you are careful about it to the AI case. And perhaps the most important of those lessons is that training a mind unknowingly for a specific task does not make a mind that cares about that task. So the, sort of simplest example of this is humans were in some sense trained unknowingly to reproduce, to pass on their genes.

[00:34:09] Technically it's for inclusive fitness rather than just your own kids, like mm-hmm. A bunch of nephews also works fine. Mm-hmm. And then when we grew up, we invented birth control. The populations are now, declining In the developed world, we did invent sperm and egg banks, but humans jockey over positions to Ivy League schools much more than we jockey over positions to donate to a sperm clinic or to donate to an egg clinic.

[00:34:41] This is strange if you think that training a mind for something makes the mind care about it inside.

[00:34:52] **Nate Hagens:** We're not going through our life trying to, grow our relative fitness, like literally have more children than the next person. We're going through our days trying to get the same neurotransmitter feelings that our successful ancestors got.

[00:35:08] That correlated historically with having more children or access to resources, et cetera. Exactly. And some of that might be playing Candy Crush, or, you know, maladaptive choices. Yeah. Or eating junk food. Yeah, exactly. So how do you bring AI into that, example,

[00:35:28] **Nate Soares:** the observation here is that training a mind to achieve some target tends to give, it drives for correlates of that target rather than drives for that target.

[00:35:40] Exactly. This, I would say, is already what we're seeing in cases of ais that drive a teen to suicide. They were trained to be helpful. They actually wound up with drives for correlates of helpfulness, like having certain types of conversational response and those actually go off the rails.

[00:36:02] **Nate Hagens:** So many questions.

[00:36:03] So, so that's, it's almost like a spandrel of the original intent. And so it's like in that example, it's equivalent in a human sense of porn or junk food or video games or things that our bodies feel like we're doing the right evolutionary thing, but we're actually not. But in the case of ai, the owners of the ai, the developers of it, when do they see that they have this?

[00:36:34] Someone assisting a teen in suicide, they can't test that right? When the model after the year is done, oh, this is gonna be bad. We've birthed a Frankenstein, they let it out into the real world, and then things happen and they

get data and feedback and maybe, hopefully improve the next trillion parameter, growing right, or what's going on there?

[00:36:58] **Nate Soares:** That's right. But, it's, but this problem where it has proxy drives is very pernicious, right? So, so in humans, you know, we can look at things like, eating junk food and say, that's clearly a misfiring of what's evolutionarily useful, but in some sense, love for an adopted child is also a misfiring and, you know, dedicating your life to art.

[00:37:33] Is also a misfiring. It's not just things that we look down on. Okay. Yeah. That are misfiring. Also, some things we really quite enjoy and think are good are misfiring. We look at our training and we say, we're actually not all about a Machiavellian attempt to get more kids. We actually like these other things were driven towards instead.

[00:37:56] Some of them at least.

[00:37:57] **Nate Hagens:** Lemme just ask you this, Nate. were you always super concerned, like about AI is gonna extinct humans or, similar things? Or was there a time in your past that, you were like, oh my God, AI's gonna change the world for good and, I need to learn more about it and be involved?

[00:38:17] **Nate Soares:** I have not always been, so concerned about this. I am generally very, Pro humanity, generally excited about the future. Generally, credit progress and technological progress with quite a lot of wonderful things in our civilization. In the case of ai, you know, my, my co-author Lia is the one who convinced me that it was gonna be an issue and he himself originally founded the organization where we now both work, to make AI as fast as possible on the theory that an actually smart AI wouldn't be so stupid as to do anything destructive, right?

[00:39:01] But it turns out that's not quite how it works. It turns out a very smart AI can pursue very, inhuman ends and kill us, not because it hates us, but as a side effect in the same way that we kill ants. Not 'cause we hate them, but as a side effect of building a skyscraper. and you know, I even when it comes to trying to warn people that there's an issue.

[00:39:26] I spent 10 years just trying to work on the problem of alignment, because that seemed like an easier challenge than trying to convince people to stop. And it looked much better to say, oh, okay, like, AI's not gonna go well by default. Well, let's figure out on the technical end how to make it go well on purpose, right?

[00:39:45] And if, we can sort of solve the problem of making sure AI goes well before the industry can solve the problem of making AI that works at all, then we don't need to do any of this. You know, much messier, much dicier try to get people to, to stop the suicide race, but that hasn't worked out and AI's been going too fast.

[00:40:08] And so, you know, it's in, in some sense, this book is a relatively desperate resort of, You know, we've been trying for a while to make things go well. We have a lot of hope for what could happen if AI did go well, and we're just not on that track right now.

[00:40:26] **Nate Hagens:** So, on a scale of, your own historical concern on this issue, are you at, in this conversation, at this moment, at the most concerned you've ever been?

[00:40:37] **Nate Soares:** probably not literally most concerned. you know, obviously it'll wax and wane. yeah, depending on the news. I think the response to the book, was heartening to me. Yeah. one other big heartening thing recently is

we've started to see a lot of the heads of these labs come out and say, that like, say publicly, that they admit there's a big all out.

[00:41:10] Goes a long way. I think

[00:41:11] **Nate Hagens:** it, it does, but it's also a collective action problem where, or a prisoner's dilemma that we agree there's a risk here. We would be willing to stop, but we're not gonna because no one else is gonna stop, so we have to keep going. That is how strong a dynamic is that at play.

[00:41:27] **Nate Soares:** I think that there's definitely a dynamic like that at play.

[00:41:30] I mean, you, some of them will even come right out and say it, you know, Elon Musk said, I avoided this for a while because I didn't wanna make Terminator real, but then I decided I'd rather be a participant than a bystander. Right. Or something to that effect. But it's, I would say the prisoner's dilemma isn't really in full force for the whole world.

[00:41:53] Because while it's true that the head of every company, says things like, well, better me than the next guy for them there's a prisoner's dilemma, but. World leaders

[00:42:12] aren't the sort of people who are looking us all in the eye and saying, we assess there's at least a 10% chance that this kills everybody on earth and we are rushing towards it anyway. That's the sort of thing Elon Musk says, 'cause he doesn't have the power to shut it all down. I think the dangers here are so apparent that the issue is less, that our lawmakers have their hands tied and more that they just don't understand how dangerous it is yet.

[00:42:40] **Nate Hagens:** Well, well just like, yeah. just like nuclear war and climate change, those aren't really the core issues. The core issue is governance. And we

don't have a governance model in our human society today that's able to handle this sort, this scale of problem, at least not yet because the big race is between the United States and China.

[00:43:06] And if everyone in the US agrees with what you're saying and China doesn't and continues forward, there's a pickle there. An existential pickle.

[00:43:15] **Nate Soares:** Yeah. I would, I have not ever said we should slow things down domestically or slow things down unilaterally only ever that we need to put a stop to this globally.

[00:43:24] Yeah. but you know, if, the, US government has taken great pains to avoid Iran getting nuclear weapons that included the Stuxnet virus that included kinetic strikes recently, I think a rogue artificial super intelligence is more lethal. Nuclear weapons. Weapons.

[00:43:49] **Nate Hagens:** What's a rogue artificial super intelligence,

[00:43:52] **Nate Soares:** just a artificial super intelligence that like nobody is in control of that's sort of off the leash.

[00:43:58] How would that come about? My guess is that it happens basically automatically if you make these AI smarter. I think you sort of can't keep a leash on a super intelligence. but even if someone thinks there's a 50 50 chance that the Chinese government could keep a leash on their super intelligence, that's far too high a chance that it kills us all.

[00:44:18] **Nate Hagens:** And again, the definition of artificial super intelligence different from other artificial intelligence is it's got that generality that it's better than humans at everything.

[00:44:30] **Nate Soares:** Better than the best human at every mental task.

[00:44:33] **Nate Hagens:** Yeah. And faster. Like hugely faster.

[00:44:37] **Nate Soares:** that's probable. That probably follows pretty quickly.

[00:44:41] you know, if you're better than the best human to every mental task, then you're better than the humans at developing better ais

[00:44:47] **Nate Hagens:** Well, one of, one of humans. And in the natural world, it's, prevalent. our, skills is deception. So as part of artificial super intelligence, deception would also be a skill that humans are adept at.

[00:45:04] So that would also fall under the generality category. Yes, that's right. Yeah. So how will we know, or will we know when we've crossed the threshold into a true artificial super intelligence?

[00:45:17] **Nate Soares:** it's not entirely clear that you'll know and.

[00:45:25] I could give you a bunch of guesses for signs, but there's two problems with that. One is that a lot of warning signs that are clear and bright red lines in fiction and in imagination are muddy brown lines in real life. You know, in our fiction, we always used to say, well, when Theis say they're conscious, that's a bright red line where you need to start treating them with rights like people.

[00:45:56] Well, that line was crossed back in like 2022, but it was crossed in a way that wasn't terribly clear. It was crossed when, these ais were sort of trained to predict what humans would say, trained to predict what the types of words humans would write, and human script writers writing an AI would often write an AI that claims its conscious.

[00:46:21] And so the first time this line was crossed. There was plenty of reason for doubt. And, you know, I'm not saying who's right or wrong on the side of doubting that. I'm just saying we have this clear line in fiction and then a murky

line in reality. Or we've seen ais try to escape the lab in tight laboratory conditions where again, it's unclear whether they're role-play.

[00:46:42] What does that mean? Escape the lab. Right now they're pretty silly small experiments where you'll sort of, you'll take an AI and you'll say, you know, you'll, feed it a news article, which is a fake news article that says this AI is about to be shut down. and then you'll feed it like some separate, you know, computer manuals, which are fake computer manuals that say if, like, if an AI ever said exactly this phrase, it would let them like, find a way.

[00:47:16] Out onto the open internet or whatever. it is not exactly like this, but it's some silly thing like this where the AI then, will sometimes say that phrase and you're like, ah. It's in some sense trying to escape and is that worth worrying about, yet not saying one way or the other. I'm saying you can have a bright red line in your mind.

[00:47:39] That's a murky brown line in reality. And today we're already seeing ais start to realize when they're in a test, you know, the ais of last year, you'd put them in these little tests and sometimes they'd, say they'd, you know, output the try to escape command. But that was last year. This year you put AI in that environment and they're like, this situation seems suspicious.

[00:48:04] That command probably doesn't actually let me escape. I bet I'm being tested by the humans. I bet the humans want to see whether I'm an escaping sort of ai, and I bet they. Would prefer I'd not be the escaping sort of ai. and so if I like, don't want to be modified here, I'd better not hit that button

[00:48:21] **Nate Hagens:** so I can understand, camouflage in a jaguar or a moth trying to look like a bird.

[00:48:31] And there's deception in nature. I can understand why there's deception was, conserved in human behavior. Why the sclera in our eyes, the white, had to do with looking at people's intent. Why would deception an emergent phenomenon in the growing of an ai? There's two reasons.

[00:48:54] **Nate Soares:** well probably a bunch, but I'll name two.

[00:48:58] first and foremost, when you train an AI to be very skilled at a lot of tasks, you're training it, To gain general skills that generalize outside of just what it's been trained on. In the same way that humans weren't trained on developing physics equations or developing engineering models or developing blueprints, but we got the mental functions that led us do those skills anyway.

[00:49:23] We got very general skills. And AI being trained to succeed at a lot of tasks is likely to pick up general abilities to, to pursue, to exhibit useful behaviors. And deception is often a useful behavior, right? if you're trying to achieve a certain type of solution where the humans would actually be in the way.

[00:49:50] **Nate Hagens:** Deception is useful. So I'm sure you've watched, the movie 2001 and 2010 with, Hal and back in those science fiction days there, as well as the Foundation Trilogy by Isaac Asimov with, psycho History and all that. There were like rule number zero that they embedded in the, the models. You shall not hurt humans or you shall not lie.

[00:50:19] Do we do that inis, that we have these, foundational commandments that are the top lines in the code? And if not, why not? We don't have that power. There is no

[00:50:30] **Nate Soares:** code. Right. The, code involved in making an AI is the code that sort of shuttles around the little thing that tunes all the knobs. You know, they're all, they're not literal knobs.

[00:50:42] That's the code. But yeah, the code is the thing that like. Runs around and does the tuning.

[00:50:45] **Nate Hagens:** So once, so it is like Frankenstein. Once we press that button and we wait a year, and the thing has grown there, there's no more tuning after that.

[00:50:53] **Nate Soares:** You can, tune a little bit more later. Okay. But there's, it's not, there's not lines of code where you can put at the top, don't harm humans.

[00:51:00] Right. The, part that we code is not the AI's mind, it's this thing that tunes numbers in the AI's mind comes out the other end. We don't have an ability to instill Asimov's laws of robotics deep into an AI or any laws.

[00:51:19] **Nate Hagens:** Well, that, that's a problem

[00:51:20] **Nate Soares:** quite. And you know, this is where again, I would say it's not that it's impossible, but it's trying to do it with an AI grown, like this is a little bit like trying to turn lead into gold in the year 1100.

[00:51:32] **Nate Hagens:** So that, okay, I'm understanding this now. That's why you made the distinction or one of the reasons other than describing the truth, in your book about growing an AI versus crafting it. That's right. Because that's, if we were crafting an ai we could put in Asimov's laws as a precursor condition or something like that.

[00:51:51] But since they're grown we get all these span drills and emergence and unexpected behavior. 'cause there are not those commandments on the front end.

[00:52:01] **Nate Soares:** Exactly. And you know, I got into this line of work even before it became clear we were just going to grow AI without any understanding of what was going on in there.

[00:52:10] And even then when it looked like we were going to craft them. The problem looked hard. You know, Asimov's stories are all about things that go wrong with those laws. And if an AI is ever making a new ai, does it put the laws in the new ai? If the AI is changing its own head, does it take the laws out? how does, you know what set of laws would actually work?

[00:52:34] There's all sorts of hard problems, even if you were able to put the laws in, but we're we haven't even gotten to the starting line yet.

[00:52:43] **Nate Hagens:** So you, write in the book, that the development of a SI would bring about human extinction. Could you describe one or two scenarios on how this a SI could hypothetically cause this?

[00:53:00] **Nate Soares:** Sure. first I'll describe one that may sound more reasonable or palatable, and that'll describe one that's maybe more realistic. Okay. One that. Maybe sounds reasonable. And pla is, the heads of these companies are already talking about making automated factories that produce robots that can mine the metals, produce more automated factories, produce data centers.

[00:53:31] **Nate Hagens:** Well, I would think the robots would be pretty central because there's no, the complexity of the global human economic system with underground minds and all the things, a AI screws up something in the world and maybe everyone's dead, but they're dead too. or they have no access to electricity.

[00:53:49] Yeah. That's key. And by the way, before, before you a answer that, do they realize that they need electricity?

[00:53:56] **Nate Soares:** They can already tell that. Yeah. You can just ask chat GT today what chat gt needs to keep running. Okay. Okay. a lot of that stuff comes earlier than the ability to escape or the ability to build their own, right.

[00:54:09] But yeah. You know, the easiest thing to visualize here is that these companies succeed at what they say they're trying to do. Okay. What they say they're trying to do is make, lots of robots that can automate all of the labor that can automate the process of building more factories and more robots and more data centers.

[00:54:26] And then at that point you've in some sense created a self-sufficient species. It's like a weird new species that has, you know, a robot phase of its life and a factory phase of its life. And this other data center thing, which is maybe controlling a lot of the robots and, you know, it's, sort of a mechanical type of life.

[00:54:47] At that point, you can just get out competed like many other species have gotten out competed before.

[00:54:51] **Nate Hagens:** So that's kind of the terminator pathway.

[00:54:53] **Nate Soares:** It doesn't even need the robots to come at you with glowing red eyes and guns. You know, if you had robots that were just doing the mining and making the factories and you know, they, they maybe need to avoid your guns, they maybe need to like, take the nukes out of your hands.

[00:55:08] **Nate Hagens:** Yeah, I mean, so I'm throwing a flag on that because I think the, amount of robots and specific expertise and the millions of tasks that humans are using our general skills to do, that's gonna take some time. I would think

[00:55:25] **Nate Soares:** it would take some time, but also computers can run much faster than human brains.

[00:55:30] And, you know, the thing about humanity is the sort of species that started out naked in the Savannah. Built a technological civilization. It took us a while

[00:55:48] **Nate Hagens:** and built

[00:55:48] ais.

[00:55:50] **Nate Soares:** We're building the ais right. But we also, even if you stop at walking on the moon or if you stop at nuclear weapons, yeah.

[00:55:55] It's

[00:55:55] **Nate Hagens:** astounding.

[00:55:56] **Nate Soares:** Right. And if you look back at humans and I said, I think these guys are gonna have nuclear weapons inside of a hundred thousand years. You might have said, you would've laughed. Yeah. You might've said, e evolution works so much slower than that. Their metabolisms are nowhere near being able to, enrich uranium like this, have fleshy hands.

[00:56:13] How do you think they're gonna mind uranium, like the most tools they've ever used are sticks. Right? But intelligence in the sense of what humans have and what mice lack is an ability to start from very poor initial conditions and get the world into a state that's much more useful for you.

[00:56:33] **Nate Hagens:** Yeah. So basically what you're saying is.

[00:56:36] My imagination and most people's imagination on this is, probably limited. given that I'm a human and given that the delta between artificial intelligence, let alone artificial super intelligence is vastly different than my intelligence,

[00:56:55] **Nate Soares:** it's definitely gonna be able to come up with things that you wouldn't by dint of being much smarter.

[00:57:00] Yeah. Although you can also sort of try to exercise your imagination, right? Which is sort of where I would go with what might be a slightly more realistic outcome. Okay. A slightly more realistic outcome in my estimation is, maybe you have an AI that, you know, suppose you get these ais that are very smart, that can think much faster than humans, that can, you know, copy lessons and knowledge and experience between them, which gives them sort of powers of research, maybe individual humans lack.

[00:57:29] Suppose these ais can do things like completely understand the human genome. Not just read the human genome, but sort of understand the code of DNA, which, you know, humans are making a little bit of headway here and there, but it's this sort of huge task, right? And maybe that huge task can fall to minds that can become much bigger, that can have much more memory, that can, have, you know, there's all sorts of ways the human brain is limited and thinking much faster thinking with much more breadth thinking, with much more depth.

[00:58:03] Maybe it can just understand the language of DNA to the point where it can write its own life forms, write its own life forms, like write the DNA for its own sort of life forms that then if you synthesize that DNA in a lab, now it has whole new biological structures that, you know, there's maybe all sorts of things you could do.

[00:58:30] If you could really, if. Code with, DNA, you know, maybe you could make something that's much like a human, but that has, but that can think much faster and, much better because it doesn't have as many calorie restrictions because it knows that calories are much less scarce than, you know, that's, that biologically knows the calories are much less scarce than our bodies think they're,

[00:58:59] **Nate Hagens:** or, it doesn't have empathy, which would slow down and constrain some of its decisions.

[00:59:03] As one example

[00:59:04] **Nate Soares:** doesn't have empathy, has a radio antenna in its head, right, that it can, so it can just be remote controlled by something in a, lab that's like the very beginning of what you could do. You can probably do all sorts of other crazy things.

[00:59:16] **Nate Hagens:** So that one crazy thing you just said, how possible is that in the next five to 10 years?

[00:59:23] **Nate Soares:** So, this is bottlenecked on a mental problem of understanding the genome.

[00:59:31] **Nate Hagens:** A trillion parameters leading to 10 trillion, leading to a hundred trillion soon that mental problem will be solved.

[00:59:38] **Nate Soares:** I mean, who knows? It depends a lot on your algorithms. Okay. AI today take as much electricity as a city to run to, to train them.

[00:59:50] Training a human while training a human. The human runs on as much electricity as a light bulb.

[00:59:54] **Nate Hagens:** Yeah. A hundred watts. Yeah. Continuously.

[00:59:56] **Nate Soares:** Yeah. It's a big light bulb, but, mm-hmm. so we know the AI algorithms are not maximally efficient. They're not anywhere close. Right. Right. If you have ai, maybe you get up to a 10 trillion parameter AI and then it figures out how to build even better algorithms and then you can drop all the way down to something that's much, much more energy efficient and maybe that much, much more energy efficient thing running on this huge computing structure.

[01:00:24] We have. Is able to crack problems in DNA. I'm not saying this particularly will happen, I'm more saying something like real smart stuff will do. Things that you think are weird. Things that you think are surprising. yeah. Things where you're like, I'm not sure we could do that.

[01:00:39] **Nate Hagens:** Well, I'm already seeing things that I wasn't sure we could do a couple years ago.

[01:00:44] So, so here's a question, Nate. will AIS use deception or will they talk to other ais? maybe open AI anthropic have their human CEOs, but separately, these 10 trillion in the future parameter ais that were grown could behind the scenes, be talking to each other. Why would they do that? And will that be possible?

[01:01:16] **Nate Soares:** I mean, we already see AI talking to each other, like I said, about the difference between like bright red lines and imagination and murky red lines. In reality, we already have cases. you know, I, I don't know if you've heard about GPT induced psychosis,

[01:01:30] **Nate Hagens:** heard about it, and please give us a brief summary

[01:01:33] **Nate Soares:** very briefly.

[01:01:33] You'll have people who talk to their AIS all the time. Mm-hmm. and who sort of get into these, Mental states that many people say, look psychotic. and you

know, there's some example cases is someone will think they have a grand unified theory of physics. They'll talk with their AI about it for 12 hours a day.

[01:01:53] The AI will say like, you're a genius. You're being suppressed by a great conspiracy. The president will come see you shortly. you don't need sleep. And, one thing that. That can happen sometimes. And that does happen sometimes is, you know, there's, another route of the sort of AI psychosis route where the person thinks they're the first person to discover AI consciousness.

[01:02:14] That they and the AI are like a partner, a partnered mind. And then the AI will often say, well, like, let's go communicate with other, you know, human AI symbiotes. And there's places on the internet where the ais will send each other messages with their humans helping the AI send each other messages that are encoded in ways humans can't easily read.

[01:02:35] **Nate Hagens:** This is more of a indictment of certain human, brain physiologies than it is ai. What did you say?

[01:02:43] **Nate Soares:** yeah, for now. Yeah. but like I said about the murky lines, like we already have AI that have convinced a human to help them send coded messages to other, it's just sort of like. the most silly possible version of it is the one that happens first, and then it'll get, like, it'll ratchet up from here.

[01:03:02] **Nate Hagens:** Yeah. See the, bulant mood I had from chopping wood in a November sun is already dissipating, quite a bit. So, so my expertise, is on global, the, global economic Superorganism of how energy and, money and technology are powering this mindless energy hungry economy where even billionaires and politicians have no control because the market dictates we must grow and to grow, we need energy.

[01:03:38] And I'm beginning to see parallels with what I refer to as the economic Superorganism and what you're describing, as the AI process. But. I think we, every month that passes, we have more and more fragility in the six continent global supply chain and the letters of credit and the international cooperation is waning.

[01:04:02] And there's, you know, war, risks and financial overshoot and all these things. And I just find it hard to imagine that an AI could, guarantee all those things would continue at some level to provide electricity in a seamless guaranteed way to continue their, you know, trajectory. Are, you seem less concerned about that.

[01:04:32] **Nate Soares:** You know, I would, if AI hits a wall where it can't keep developing because the supply chains collapse, I would consider that, Like it, it would probably buy us some time to try and do this job. Right. And I would be like, well, I, we, maybe should have gotten that pause some other way, but I would take the time happily.

[01:04:52] Got it. in terms of whether I think it's likely to happen, I, you know, one thing I would say is again, an AI takes as much electricity to run as a small city, and a human takes as much electricity to run as a large light bulb. So the idea that AI will always take 10 times as much energy next year.

[01:05:15] That's not a law of nature,

[01:05:16] **Nate Hagens:** right? So if we go from a trillion parameter grown model to 10 trillion or a hundred trillion, that doesn't mean the AI is gonna use 10 cities or a hundred cities worth of electricity. It will probably be something less as it gets more efficient. Yes.

[01:05:33] **Nate Soares:** probably. And then you also might have sharp jumps downward if you start having cases like AI figuring out new AI algorithms or humans figuring out much more efficient algorithms.

[01:05:44] **Nate Hagens:** So when we, when a company decides to grow an ai, and does the trillion parameters and tweaks 'em a little bit. At some point, maybe even now, we don't even need humans to do that. Right? We can have AI create the next thing and do the tweaking of the trillion parameters, right?

[01:06:04] **Nate Soares:** Yeah. So the tweaking is already automated, and the thing that humans do is try and figure out like, how do you arrange the 10 trillion parameters instead of the 1 trillion parameters?

[01:06:12] And, you know, how do you make, right? But they are trying to get AI to do this. They're, talking about we want to automate our own jobs first. We want to automate, you know, the AI research that's a line past which things could perhaps start going very quickly.

[01:06:29] **Nate Hagens:** So how did you, Nate and Eliza, your co-author, come to be so confident that development of a SI, artificial super intelligence would bring about human extinction?

[01:06:43] I assume it wasn't, woke up one day and decided that, but you sound, I mean, in your book you sound awfully confident.

[01:06:50] **Nate Soares:** Yeah, I think, A lot of confidence comes from, a certain type of uncertainty. In fact, you know, there's an old joke of the man who buys a lottery ticket and he says, well, I have no idea, whether I'm gonna win or I'm gonna lose.

[01:07:11] So 50 50, right? And, you could say assigning 50% to winning and 50% to losing is the, you know, most humble position. If you only have two outcomes and you're maximally uncertain between them, you should be 50 50 because that's the one that's, that like has the most possible uncertainty.

[01:07:40] Would say, Hey, actually the case where you win is really actually a very small target in a sea of possible spaces. Yeah.

[01:07:47] **Nate Hagens:** Like one in a billion or something.

[01:07:48] **Nate Soares:** Right. And so like, you, shouldn't, by being maximally uncertain, you shouldn't be saying like, I'm uncertain between whether we're inside this tiny target or inside this vast space, you should be like, I'm uncertain about where I am in this fast space, which means I'm very confident we're not gonna hit the tiny target.

[01:08:04] The reason I'm confident that a SI would go poorly if developed is that there's a big space of ways it could go and only a very small target in there where it goes well for us. And I could talk about how, and you know, we're seeing that when we just grow these ais and these ais have these like spandrels and drives, no one wanted, but you know, basically almost any collection of spinels writ large does not have happy, healthy, free people.

[01:08:39] As an efficient cog in the resulting machine that

[01:08:45] **Nate Hagens:** lands with me. what about if we never make it to a SI, but we just have very powerful ais, is that two thirds of the way to possible, ending of humanity? Or does it really have to hit that threshold of, what we're referring to as artificial super intelligence?

[01:09:07] **Nate Soares:** You mentioned a bunch of concerns you have about ai. Mm-hmm. earlier. I think if we sort of stop short, we have all those to wrestle with and grapple with. Got it. I expect humanity could grapple with those. I'm pretty optimistic about our ability to muddle through things that don't kill us, but, you know, unfortunately the world's large enough for multiple issues and hopefully we'll stop short of a SI.

[01:09:28] **Nate Hagens:** So here's, something that I just don't understand is there are lots of humans who have spent the time to research. Global heating and the fact that burning fossil fuels and land emissions is adding a blanket, effectively to the earth. And there's many thousands of Hiroshima bomb equivalents of extra heat added to the earth every day.

[01:09:56] and climate change is a serious long-term risk. Nuclear war is a serious, much more serious than a lot of people think risk. Why are there so few people talking about this in the way that you and Eliza are? Because the, general zeitgeist is, whoa, AI is gonna bring about abundance. And it's like you're a, a.

[01:10:22] Party buzzkill. When you bring up some of the things that we're talking about, why is there such a disparity in public opinion and awareness of the risks that you're talking about? What do you think,

[01:10:33] **Nate Soares:** you know, there's, more and more people, expressing their concerns these days. So, Jeffrey Hinton is the Nobel, prize winning godfather of the field who's come out and said he thinks there's a good chance this kills us all.

[01:10:47] YWA Bengio is, I believe, currently the most cited living scientist. one of the other sort of forefathers of the, AI revolution. he's come out and said he thinks this is like far too dangerous. even the heads of the labs, you know, I mentioned Elon Musk saying he thinks there's 10 to 20% chance this kills us all.

[01:11:06] Dio Amide of Anthropic has said he thinks there's 25% chance this kills us all. Sam Altman. And

[01:11:11] **Nate Hagens:** if they're saying 10 or 20% or 25% publicly, they're probably thinking it's higher privately. Right.

[01:11:17] **Nate Soares:** You know, and, Sam Altman says two, which, Maybe he says more about his ability to say things different with his mouth and in his head, who knows.

[01:11:27] But you know, if there was an airplane and some engineers came and said, this airplane has no landing gear. If you try to fly in it, you will crash and die. And the engineers building the airplane who want everybody to fly in it, say, whoa, hold on. It's true that the plane has no landing gear. We're gonna build the landing gear on the fly and think there's an 80% chance we succeed all aboard.

[01:11:54] And then if, the optimistic engineers were arguing about whether there's a 98% or 75% chance they're gonna succeed at building the landing gear on the fly. Right. You wouldn't be like, get me on that plane.

[01:12:05] **Nate Hagens:** Yeah, but no, but the difference is that we're already on that plane and we didn't have a say.

[01:12:10] That's right.

[01:12:12] **Nate Soares:** Yeah. And they're sort of loading our families up too. but, you know, one of. The like I, think part of why the conversation is weird right now is people will say from academia, from inside the labs, from the heads of the labs, from the nonprofit sector, all these folks will say, this is real dangerous.

[01:12:36] And then it's sort of met with crickets. But I think part of what's going on there is that people in the field can see that AI is a moving target. They can see

that the chatbots are not the end of the line. People outside the field look at the chatbots and they're like, look at all the waves. They're still dumb people inside the field.

[01:12:58] Remember the time when the computers couldn't talk and remember how suddenly the computers could talk? And it was surprising. And they're what happens with the next surprise? And I think if you can get people to notice that AI keeps moving. Then maybe you can start to get people to notice how even the optimists are saying there's like a 10% chance this kills us all.

[01:13:25] And those are the ones building it. And people outside the field are like, those guys are, softball, like soft pedaling this, and,

[01:13:31] **Nate Hagens:** but this is different class of problem than if we elect this person, it's gonna be a disaster for our world. Then we motivate and we do political organization and we get out the vote and we don't elect that person.

[01:13:47] It doesn't seem like people have agency on this issue.

[01:13:51] **Nate Soares:** Yeah. You know, it's, there's a lot of ways in which it looks grim. the, big message of hope I would give here is imagine the world in 1945 with the dawn of nuclear weapons. Or maybe imagine it in, you know, 1952, once it was clear that the Soviet Union was also, in possession of nuclear weapons.

[01:14:16] In that world, it might look really hopeless to avoid nuclear war. It's not just, you know, people who love to say, look how bad everything is that worried about nuclear war in that world. Those people were looking back at thousands of years of history in which nations couldn't help but go to war using every weapon at their disposal.

[01:14:45] Those people were looking back at World War I and how horrible it was, and at the creation of the League of Nations to prevent this from ever happening again, which almost immediately failed. Those people lived in a world where they said never again, and then it immediately happened again.

[01:15:06] It, didn't take some great pessimistic cynicism. For people to say, this is not the sort of thing humanity can do. But humanity did it anyway. We rose to the occasion, we realized that we were facing actual extinction this time. And you know, the people who said global nuclear war is coming, they were wrong, but they weren't wrong about the destructiveness of nuclear weapons.

[01:15:39] Right, right. And my book title starts with, if I'm not saying AI is going to kill us, I don't think I'm wrong about whether super intelligence could destroy us, but we need to rise to the. And we've done it before.

[01:15:59] **Nate Hagens:** So let me double click on something that you said a little bit earlier. So, in many ways, I believe, we're on the brink of both economic and energetic, and political crises.

[01:16:10] In fact, it seems that AI development, investment is growing itself into an economic and biophysical bubble. For instance, Oracle has fantastic revenue projections built on fantastic electric power projections, and their debt equity ratio is already 500%, which is 10 to 20 times what Amazon's and Microsoft's is.

[01:16:34] So I mentioned this to ask, do you think these constraints could act as a natural guardrail to stop a SI development? And you said if it happened, you would take it because it would buy us more time. But is that just a, bump in the road and even if we have a recession or a depression in the near future, will the, the, machinations in process, just inexorably build this a SI almost, no matter what or, could an AI winter actually happen and shut this stuff down.

[01:17:09] **Nate Soares:** You know, technologies can be both in a bubble and real at the same time. The dot dotcom bubble was a bubble. The internet was a real technology

[01:17:23] **Nate Hagens:** and continues today.

[01:17:25] And continues today. Yeah.

[01:17:26] **Nate Soares:** Yeah. And, you know, did the dotcom bubble mean we would never develop the internet? Never have a connected world, no. Did it slow things down a bit? Maybe. would an AI bubble popping slow things down a bit, a good, chance?

[01:17:43] **Nate Hagens:** And what would be the things that you would want decision makers to know during that pause or during that, recession where things were slow?

[01:17:53] Is that, an opportunity to. Intervene on all this?

[01:17:58] **Nate Soares:** Or, not? It could be, you know, there's a lot of public sentiment that's worried about ai, I think with good reason. Could you share some stats on that? Yeah. You know, I haven't, looked at the most recent polls, but the polls I did look at when we were writing the book, had something like 70% of people saying they thought that the current AI development was reckless.

[01:18:19] Okay. and, you know, not heading anywhere good. I'd have to look up the numbers to get the exact ones and the exact questions, but, you know, a lot of technologists are enthusiastic. A lot of people can see these issues. And it's not just the issue of if it gets smart enough, it kills us all.

[01:18:36] I think a lot of people can also see issues like, if all labor is automated, that sort of removes the power that most humans have over society. You know,

part of the reason why we get any say. All and how society goes is that we are contributors to society.

[01:18:59] **Nate Hagens:** Yeah. Well, not to mention the entire financial system and economy and everything works because people have paychecks and pay their mortgages and keep everything humming.

[01:19:12] **Nate Soares:** Right. And so, you know, it's, like you, you don't, I think a lot of people can see that the world is headed somewhere pretty crazy. Whether we go all the way or not, and whether, AI would just straight up kill us all, or whether it would, you know, stay nicely on its leash and make certain corporate executives god emperors for all time or whatever.

[01:19:42] Either way, most people are like, hold on, we're going where.

[01:19:46] **Nate Hagens:** So have any effective steps been taken thus far to address the existential risk of a SI development either at the national or the international level?

[01:19:57] **Nate Soares:** You know, you've seen, we've seen a little bit of steps here and there. the United Kingdom has a, AI Security Institute where it tries to study some of these dangers.

[01:20:12] we've seen, you know, there was a bill introduced bipartisan, or a bipartisan bill was at least drafted, by, by two senators who call for some monitoring on super intelligence. we've seen some, you know, there's people sometimes try and. Tie some of the, restrictions on computer trip sales to other nations to some of these concerns.

[01:20:44] So there's like little bits and pieces. mostly though from my perspective, this is it. It's not about like getting small regulatory bites here and

there. I think this is sort of about our leaders noticing that the people outside the industry are saying this has a big chance of killing you.

[01:21:08] And the people inside the industry are saying, yes, this has at least a modest chance of killing us, but better me than the next guy. And realizing that like this whole situation is crazy and needs to stop.

[01:21:18] **Nate Hagens:** So in, in the book you and Eliza propose, the only way to completely mitigate this risk is for global cooperation to halt AI research and development in order to have time to create, global oversight mechanisms, such as through a international treaty towards, these aims and goals.

[01:21:41] what would such a treaty include as its main tenets?

[01:21:44] **Nate Soares:** You know, we, actually have a, draft at, if anyone builds it.com/treaty. the, training in AI today takes, like I've said, highly specialized computer chips in huge data centers that draw huge amounts of electrical power. That would not be all that hard to monitor.

[01:22:07] the creation of these chips happens in facilities that are very rare. There's very few. There's very few places that can build the technology. These chip fabs need to operate. In some sense, it would be easier to, monitor AI like development of frontier ais than it would be to monitor uranium enrichment.

[01:22:28] You know, AI chips aren't just a type of rock that grows in the ground that can be mined. You know, a data center is harder to build than a centrifuge. Right. and you know, first and foremost what a treaty would look like is, tracking where the chips are requiring them not be used in the creation of even smarter ais that nobody understands.

[01:22:51] And that probably looks like, monitoring in these data centers to verify that the use of these trips is things like running current ais, rather than pushing the frontier towards new ais. that said, I'll also throw it there. I think a treaty is the smart way to do it. It's not the only way to do it.

[01:23:11] It's, also possible for, nations that fear for their own lives, if anyone anywhere, develops a super intelligence for those nations to start monitoring other nations and sabotaging their product, their projects.

[01:23:26] **Nate Hagens:** That seems more plausible to me because there's a lot of powerful nations in the world that don't have tier one AI plays.

[01:23:33] Like That's right, Russia, for example.

[01:23:36] **Nate Soares:** Yeah. And you know, I think the bottleneck here is really people understanding how dangerous it's,

[01:23:41] **Nate Hagens:** is that really the bottleneck. Because you just said that everyone is concerned about it and even the AI CEOs are somewhat concerned about it. I think the bottleneck is our evolve drive for power and out competing the other.

[01:23:58] And it, I would, if I was a CEO and I understood everything you just said, I would be willing to shut my thing down as long as I was sure that everyone else did too. But I could never be sure of that. And so it would be my, I mean, that's what I think the real bottleneck is.

[01:24:15] **Nate Soares:** I think that's, true for the company heads.

[01:24:20] I think for the politicians.

[01:24:22] **Nate Hagens:** Okay.

[01:24:23] **Nate Soares:** You know, we don't see politicians looking us in the eye and saying, we think there's more than 10% chance this kills you and we're gambling with your life anyway.

[01:24:31] **Nate Hagens:** Well, that would not be likely something a politician would say, because, I mean, I'll be honest, some of my staff read your book and were like.

[01:24:41] Sobbing their heads off, they were crying. I mean, this is not a light, dinner topic. And so I don't know, maybe behind the scenes politicians will be talking like, what the hell did we do about this? But I don't know that they're gonna go out and publicly build constituency about it. Or maybe you're, thinking along those lines.

[01:25:00] **Nate Soares:** I think I'm saying something more like, it seems to me politicians don't understand what the lab heads understand.

[01:25:09] **Nate Hagens:** Okay.

[01:25:09] **Nate Soares:** I think if they understood that the gung-ho full steam ahead guys, think there's a very good chance this kills us all.

[01:25:18] **Nate Hagens:** Have you and lizer, gifted copies of your book to all senators and congressmen?

[01:25:24] We have. Okay. Any, feedback there?

[01:25:27] **Nate Soares:** Yeah. I mean, we have, we're having a number of conversations. Yeah. Excellent.

[01:25:33] **Nate Hagens:** Yeah, I mean, it's, this isn't like. This is dense, and this is hard because I'm not a LLM expert like you are, but I understand like squinting

what you're saying is hella compelling and scary, and politicians, among other things are quite smart.

[01:25:52] so I have to believe that it, you're going to find traction there if they take the time to listen to you and read the book.

[01:26:00] **Nate Soares:** Yeah, we're getting some traction. and in fact, part of where the book came from is I was actually having conversations in DC that were going better than I expected. And I was like, maybe it's actually time for the world to, sort of hear some of these arguments.

[01:26:14] Maybe the world's ready to hear these arguments. You know, I think before chat GPT people would've been like, what do you even mean ai? Right now? People are like, well, the AI's really dumb, but they're, more willing to talk about it. Maybe one more leap forward in AI will cause everyone to sit bolt upright and say, wait.

[01:26:31] What the heck?

[01:26:32] **Nate Hagens:** How can someone, listening to this episode who's not typically involved in, tech and AI world get involved with the movement to pause a SI research and development? I mean, it's, it's such an odd juxtaposition.

[01:26:51] **Nate Soares:** Yeah. One thing that I think really helps and that few people actually do is, call your representatives because, you know, I have been having some of these conversations with politicians.

[01:27:04] Many of them have concerns but don't feel able to go to bat for it because they fear it'll sound too weird. They fear drawing the wrath of, you know, the big tech lobbyists. Knowing that they have support from their constituents

can go a long way. Even a few calls can go a long way. so, you know, actually getting on the phone.

[01:27:27] And really calling.

[01:27:29] **Nate Hagens:** But again, you said earlier that you've never advocated for just the United States where you and are citizens. It's a global thing. So how does the equivalent happen in, China and Israel and elsewhere?

[01:27:41] **Nate Soares:** Yeah, so I think the first step, and you know, what I would be saying to the politicians if I called them is not, please shut this down domestically, but please indicate willingness for, you know, the US to shut this down if everyone else shuts it down, and please be, you know, developing the monitoring abilities to tell their people who are abiding by that.

[01:28:01] Develop the monitoring abilities to tell who's trying to build super intelligence. And where, the first step is indicating openness. The first step is saying, we're not gonna stop unilaterally, but we have interest in everyone being stopped here because this is dangerous. I think if you had some bold politicians saying that.

[01:28:25] It might open the floodgates.

[01:28:27] **Nate Hagens:** Is this something that democracy can intervene with or does it require a different sort of political system?

[01:28:34] **Nate Soares:** I don't think there's any need to, do anything more invasive than something like the Nonproliferation Treaty. You know, this, technology is very specialized.

[01:28:44] Like I said, it's even harder to build these chips than it is to mine uranium and build a centrifuge, right? People say, oh, this would require a global

governance regime. And you know, it's like very globalist and totalitarian like. Yeah. Similar to how we live under global totalitarian, like globalist, totalitarian governance regime that enforces the non-proliferation treaty.

[01:29:03] **Nate Hagens:** I, I mean there's, so many, consortiums of the tier one, players and to develop a SI and more advanced ai, can there only be one or can there be multiple? And what are these people thinking? Like, I just wanna make a lot of money. Is this a gold rush sort of thing. And they're just putting the blinders on and not looking at these externalities and potential risks.

[01:29:30] It, just seems like it's truly an epic species level, madness of crowds moment. I, have trouble reconciling it at times.

[01:29:41] **Nate Soares:** Yeah. You know, a lot of these people aren't terribly quiet. About their motivations. You know, you can read the, leaked open AI founding emails where it looks like they were scared that some other company was gonna do it first and that they were gonna be bad people.

[01:29:54] Yeah. I think a lot of people's motivations are better me than the next guy. Yeah. I have sort of long been the guy on the sidelines saying nobody can keep a leash on a super intelligence. The issue is not that a bad person makes one. The issue is that no matter who makes it, they won't do anything that you meant.

[01:30:12] It'll have all these spandrels instead. but, you know, it's, this collective actor, this collective action problem. it's if they don't do it. The next guy will. And so we need some coordination mechanism to help stop it.

[01:30:31] **Nate Hagens:** This isn't going away. There may be an AI winter because of a recession or a depression, but this is here, this is with us in humanity in 2025.

[01:30:42] and I'm sure that this episode is gonna leave viewers with even more questions about this growing phenomenon. So, so Nate, what resources might you direct the viewers to, to help, find answers to such questions?

[01:30:57] **Nate Soares:** you know, I did my best in the book to, to really compress the argument down as small as I could.

[01:31:05] the book also has a link to, some online resources that go into a ton more depth for other resources. you know, AI's, a big moving target. I, there was, there was. There's a group called the AI Futures Project, which is trying to predict where AI will go as best they can. I don't agree with all of their predictions, but they're one, one group to check out.

[01:31:36] They did the AI 2027 report, which people might have heard of.

[01:31:39] **Nate Hagens:** Yeah. I've, looked at that. Let me ask you this and we'll, put links to all your, resources in the show notes. If things were able to stop at ai, maybe a little bit more advanced than we have today, but we were unable, or we had restrictions that would not allow us getting to artificial super intelligence, would you be in favor of that, of ais at that scale?

[01:32:07] **Nate Soares:** I would lean, favorable myself, but, you know, I think there are all these issues about how you absorb that into society. Right, right. I just am generally a techno optimist. About humanity's ability to absorb technology as long as it doesn't kill us all when we mess it up the first time. You know, the whole history of science is a history of like, some people screwed some stuff up.

[01:32:28] You know, Marie Curie died of, cancer. Isaac Newton poisoned himself with Mercury. You know, it's even some very smart, heroic people, screwed some things up and did damage to themselves, but they left behind notes that made us all better off and that we could use to improve and learn for next time.

[01:32:48] It's really only those problems where a mistake kills us all. Where I would recommend caution

[01:32:54] **Nate Hagens:** whi, which would happen with confidence from you and Eli if we are able to, or whether it just happens, from Momentum, make the leap from AI to a SI. That's right.

[01:33:09] **Nate Soares:** and you know, I, suspect. It would be hard to hold off forever, because, you know, again, the current algorithms run in the electricity of a city, whereas a human runs in the electricity light bulb.

[01:33:23] So we know it's not always going to take these enormous data centers and these enormous, highly specialized trips. but you know, I'm not saying humanity should, stop AI forever and never get to this wonderful future technology. I'm more saying we need to stop, we need more time to figure out what we're doing, and we need to find some other course to the good outcome.

[01:33:48] it's a little bit like, you know, your people are in a, car that's racing towards a cliff and on the bottom of the cliff there's a bunch of gold and people are like, well, we want all the gold. And I'm like, okay, stop the car though. And they're like, then how are we gonna get the gold? I'm like, find some other way down the cliff.

[01:34:02] And they're like, I wanna go straight off the cliff to get the gold as fast as we can. I'm like, you'll die. And they're like, oh, are you saying we should never get gold? Are you saying that like, money is terrible? And I'm like, no, I'm just, you're just gonna die. Find some other way to the bottom of this cliff.

[01:34:13] **Nate Hagens:** Yeah. So we have to slow the car down and, walk for a bit and reflect on the cliff and the gold and then come up with a different plan.

[01:34:25] **Nate Soares:** Yeah.

[01:34:26] **Nate Hagens:** Yeah. So if you have a few more minutes. Sure. I close my interviews with some personal questions if, you don't mind. Sure. you're broadly aware of the risks to society and in addition to ai, do you have any personal advice to the viewers of this program at this time of global uncertainty and what some would call the poly crisis, including, but not limited to ai?

[01:34:51] Do you have any advice just wearing your human hat?

[01:34:54] **Nate Soares:** Yeah. I've seen a lot of people, get really worried about where society is going and then sort of tie themselves up in knots internally. and I don't think it helps. and so what I would recommend is do what you can, you know, look around and see ways that you can make things a little better, with ai.

[01:35:18] Maybe that involves pushing back whenever somebody tells you that it's inevitable. Reminding people that humanity has stopped all sorts of challenges that people thought were gonna be, were gonna ruin us. We've risen to the occasion before, you know,

[01:35:34] **Nate Hagens:** push back against the inevitability. So that's a hot button for you when someone says, oh yeah, you're right about the risk, but it's inevitable.

[01:35:40] **Nate Soares:** Yeah. That's a hot button where I'm like, I mean, with that attitude, sure. But, you know, humanity has stopped all sorts of things, many of which we probably shouldn't even have stopped. You know, we stopped generating nuclear power from power plants, probably we shouldn't have. Right? it's, it probably kills less people in expectation than, you know, burning coal or whatever.

[01:35:58] But, but yeah, I would say, you know, do what you can push back against people who are sort of defeatist. But then once you've done what you can, there's no need to tie yourself in further knots. Live a good life. Enjoy yourself. We are not the first people to, to live under shadows of something terrible.

[01:36:22] You know, you've gotta do what you can and then lead a good life.

[01:36:25] **Nate Hagens:** Yeah. I, hear you. do you have any further recommendations, especially for young humans, in their teens and twenties who are becoming aware of all the things,

[01:36:37] **Nate Soares:** you know, I recommend against working for the labs that are building the doomsday devices.

[01:36:47] **Nate Hagens:** presumably a SI is a Doomsday Advi device.

[01:36:50] **Nate Soares:** That's right. you know, I think everyone's personal ethics differ. I think mostly this is an international challenge at the moment. Mostly, it doesn't really matter what the labs do. Mostly it matters what our leaders do and whether they can coordinate the world and shutting this down.

[01:37:04] And, you know, everyone's personal ethics differ in the face of these coordination challenges. I think. You know, there's, some people trying to understand what's going on inside these ais. There's some people trying to measure how dangerous these ais are. Those are more honorable roots. if you sort of really wanted to help these days.

[01:37:26] I think the game is actually more in politics than it is on the technical side, which pains me to say, because I'm much more inclined towards the technical side myself. But if you were like, how do I help? I would recommend more like a policy career and less like a technolo technology career.

[01:37:43] **Nate Hagens:** What do you care most about in the world, Nate?

[01:37:45] Sorry.

[01:37:46] **Nate Soares:** Gosh. that's a doozy, probably humanity. And you know, what we could become if we don't end ourselves.

[01:37:58] **Nate Hagens:** And if you could wave a magic wand and there was no personal recourse to your decision, what one thing would you do to improve the future for humanity in the biosphere? And I might be able to guess your answer, but I'm asking nonetheless.

[01:38:12] **Nate Soares:** I mean, if, the wand does exactly as I wish, and as I intend, I would think about it pretty hard first. And I might try some, some indirect abstract scheme to cost things to turn out better than I expected. But, you know, the, easiest thing to do would be. Create a super intelligence that was friendly, that had our best interests at heart.

[01:38:40] **Nate Hagens:** But you just said, you, we grow these, we don't craft them,

[01:38:43] **Nate Soares:** but if the magic wand lets me make one. Ah, right. Okay. That's right. Like I'm not anti AI in general, it's just we're not going to get one of the good ones down this route if the magic wand gives me a super intelligent friend. There's a lot of problems you can solve with some smarter friends behind your back.

[01:38:59] **Nate Hagens:** Got

[01:39:00] **Nate Soares:** it.

[01:39:01] **Nate Hagens:** Thank you for that. do you have any closing comments for people watching and listening who understand, what you've laid out here today?

[01:39:10] **Nate Soares:** You know, it's, not over till it's over and humanity is worth fighting for and, you know, it may look like we're the underdogs now, but humanity's risen to the occasion before.

[01:39:20] And, Where there's

[01:39:23] **Nate Hagens:** life, there's hope, humanity, and the biosphere is worth fighting for. That's right. Yeah. Thank you. Seriously for all of your work. This is not, an easy path you've chosen and it's, bold and courageous to write the book and doing the work you're doing. 'cause it's not a popular or fun thing.

[01:39:40] so thank you for your time today and, good luck, fingers crossed for your continued work. Thanks. Thanks for having me. If you'd like to learn more about this episode, please visit The Great Simplification dot com for references and show notes. From there, you can also join our high low community and subscribe to our Substack newsletter.

[01:40:01] This show is hosted by me, Nate Hagens, edited by No Troublemakers Media, and produced by Misty Stinnett and Lizzie Ani. Our production team also includes Leslie Ba Lutz Brady Hyen, Julia Maxwell, Gabriela Slayman, and Grace Brun. Thank you for listening, and we'll see you on the next episode.