[Transcript] Lex Fridman Podcast / #373 – Manolis Kellis: Evolution of Human Civilization and Superintelligent AI

The following is a conversation with Manolis Gallis, his fifth time on this podcast. He's a professor at MIT and head of the MIT Computational Biology Group. He's one of the greatest living scientists in the world, but he's also a humble, kind, caring human being that I have the greatest of honors and pleasures of being able to call a friend. And now a quick few second mention of each sponsor. Check them out in the description. It's the best way to support this podcast. We've got Aidsleep, FNABS, NetSuite for business management, software, ExpressVPN for privacy and security on the interwebs, and InsightTracker for biological data. Choose wisely, my friends. Also, if you want to work with our amazing team, we're always hiring, go to lexfreedman.com slash hiring. And now onto the full ad reads. As always, no ads in the middle. I try to make this interesting. I often fail, but I try. And if you must skip them, please still check out our sponsors. I enjoy their stuff. Maybe you will too. This episode is brought to you by Aidsleep, and it's new pod 3 mattress, which I think of as a teleportation device into a land of dreams. A place where the mind goes to escape the spacetime physics of this reality of the waking world. Anything is possible in the place of dreams. The darkness that lurks in the young game shadow is possible. The hope to try and symbolize this delight at the end of the tunnel is possible. All of it is possible. It's all up to you. I'm kind of somebody that likes both the the good and the bad of dreaming. There's a cleansing aspect of a bad dream. You wake up freaking out a little bit, but then you realize how awesome this life is that whatever happened in the dream world is not real. It's a kind of dress rehearsal for a bad event that happens in reality, but it doesn't. It's like in a video game, you get to save, do a dangerous thing, screw it up, and then you get to load and try again. That's what a dream is. Anyway, I love dreaming. I love sleeping. I love naps. And an eight-sleep mattress is the best place to teleport into that dream world. Check them out and get special savings when you go to eightsleep.com slash flex. This show is also brought to you by Netsuite, an all-in-one cloud business management system. It manages financials, human resources, inventory, e-commerce, and many business-related details, all things that I have to start figuring out. I put up a job position for somebody to help me with financials. All of it needs so much help because running a business, any kind of business, whether it's a creative business or a robotics factory or any kind of AI, software company, anything you do has so many components. And I would say many of them don't involve any of the kind of cutting-edge engineering and design and brainstorming and innovation and research and all that kind of stuff. You have to do all the basic minutiae, the glue that ties together people and makes the whole thing run. And I think you should use the best tool for that job. And Netsuite is something I can definitely recommend as a great tool. You can start now with no payment or interest for six months. Go to netsuite.com slash flex to access their one-of-a-kind financing program that's netsuite.com slash flex. This show is also brought to you by ExpressVPN. My comrade, my friend, the piece of software that has accompanied me through darkness and light for many years, way before I had a podcast, way before I have found my way, though I am still forever lost. And you, if you too are forever lost, perhaps it will also warm your heart as a dead mind. First of all, practically speaking, let's put the romantic stuff aside, you should be using a VPN on the internet. And ExpressVPN is the VPN I've used and can highly, highly, highly recommend. By the way, I apologize for the coarseness of my voice. I've been feeling a little bit under the weather. Whatever the heck that expression actually means. There's always chat GPT that can ask the question, but I'm not going to. I'm just going to go with it. I'm going to wing it. Going to wing it is another funny expression, right? Wing it. What does that mean? Probably has to do with birds. And the fact that bird flight is a kind of chaotic process that's not amenable to clear dynamical system modeling that, for example, an airplane is. But let us return to the piece of software you should be using to warm your heart and to protect your privacy on the internet. Go to expressvpn.com slash Lex podd for an extra three months free. This show is also brought to you by inside tracker, a service I used to track biological data that comes from my body. There's that song. It's my party. And I'll cry if I want to. And I used to think it said, it's my body and I'll cry if I want to. I don't know what I thought that actually meant. What's a good song? It's a silly song. It's my party and I'll cry if I want to. Cry if I want to cry if I want to. You would cry too if it happened to you. Anyway, speaking of which, we're going hard on the tanges today. I like data that comes from my body that is then used in machine learning algorithms to make decisions or recommendations of what I should do with said body. Lifestyle choices, diet, maybe in the future be career advice, all kinds of stuff, dating, friends, anything, you know, but basically health stuff, medicine, I think, and that's the obvious way you should be figuring out what to do with your body is at least in large part based on the data that comes from your body and not just once but many times over and over and over and over. Inside tracker, pioneering, the data collection, sort of the blood test, then then can extract all kinds of information, give you advice. I highly recommend them. Get special savings for a limited time when you go to inside tracker.com slash Lex. This is a Lex Friedman podcast. To support it, please check out our sponsors in the description. And now, dear friends, here's Manolis Kallis. Good to see you, first of all, Manolis. Lex, I've missed you. I think you've changed the lives of so many people that I know. And it's truly like such a pleasure to be back, such a pleasure to see you grow, to sort of reach so many different aspects of your own personality. Thank you for the love. You always give me so much support and love. I just can't, I'm forever grateful for that. It's lovely to see a fellow human being who has that love, who basically does not judge people. And there's so many judgmental people out there, and it's just so nice to see these beacon of openness. So what makes me one instantiation of human irreplaceable, do you think? As we enter this increasingly capable age of increasingly capable AI, I have to ask, what do you think makes humans irreplaceable? So humans are irreplaceable

because of the baggage that we talked about. So we talked about baggage. We talked about the fact that every one of us has effectively relearned all of human civilization in their own way. So every single human has a unique set of genetic variants that they've inherited, some common, some rare, and some make us think differently. Some make us have different personalities.

They say that a parent with one child believes in genetics. A parent with multiple children understands genetics. Just how different kids are. And my three kids have dramatically different personalities ever since the beginning. So one thing that makes us unique is that every one of us has a different hardware. The second thing that makes us unique is that every one of us has a different software uploading of all of human society, all of human civilization, all of human

knowledge. We're not born knowing it. We're not like, I don't know, birds that learn how to make a nest through genetics and will make a nest even if they've never seen one. We are constantly relearning all of human civilization. So that's the second thing. And the third one that actually makes humans very different from AI is that the baggage we carry is not experiential baggage, it's also evolutionary baggage. So we have evolved through rounds of complexity. So just like ogres have layers and shrek has layers humans have layers. There's the cognitive layer, which is sort of the outer, you know, most of the latest evolutionary innovation, this enormous neocortex that we have evolved. And then there's the emotional baggage underneath that. And then there's all of
the fear and fright and flight and all of these kinds of behaviors. So AI only has a neocortex. AI doesn't have a limbic system. It doesn't have this complexity of human emotions, which make us so, I think, beautifully complex, so beautifully intertwined with our emotions, with our instincts, with our, you know, sort of gut reactions and all of that. So I think when humans are trying to suppress that aspect, the sort of quote unquote more human aspect towards a more cerebral aspect, I think we lose a lot of the creativity, we lose a lot of the, you know, freshness of humans. And I think that's quite replaceable. So we can look at the entirety of people that are alive today, maybe all humans who have ever lived and map them in this high dimensional space. And there's probably a center, a center of mass for that mapping, and a lot of us deviate in different directions. So the variety of directions in which we all deviate from that center is vast. I would like to think that the center is actually empty, that basically humans are just so diverse from each other, that there's no such thing as an average human, that every one of us has some kind of complex baggage of emotions, intellectual, you know, motivational, behavioral traits, that it's not just one sort of normal distribution, we deviate from it. There's so many dimensions that we're kind of hitting the sort of sparseness, the curse of dimensionality, where it's actually quite sparsely populated. And I don't think you have an average human being. So what makes us unique in part is the diversity, and the capacity for diversity, and the capacity of the diversity comes from the entire evolutionary history. So there's just so many ways we can vary from each other. Yeah, I would say not just the capacity, but the inevitability of diversity. Basically, it's in our hardware, we are wired differently from each other. My siblings and I are completely different, my kids from each other are completely different, my wife has she's like number two of six siblings. From a distance, they look the same, but then you get to know them, every one of them is completely different. But sufficiently the same that the differences interplay with each other. So that's the interesting thing where the diversity is functional, it's useful. So it's like we're close enough to where we notice the diversity, and it doesn't completely destroy the possibility of like effective communication and interaction. So we're still the same kind of thing. So what I said in one of our earlier podcasts is that if humans realize that we're 99.9% identical, we would basically stop fighting with each other. We are really one human species, and we are so similar to each other. And if you look at the alternative, if you look at the next thing outside humans, it's been six million years that we haven't had a relative. So it's truly extraordinary that we're kind of like this dot in outer space compared to the rest of life on Earth. When you think about evolving the rounds of complexity, can you maybe elaborate such a beautiful phrase, beautiful thought, that there's layers of complexity that make. So with software, sometimes you're like, oh,

let's like build version two from scratch. But this doesn't happen in evolution. In evolution, you layer in additional features on top of old features. So basically, when I like every single time, myself divide, I'm a yeast, like I'm a unicellular organism. And then cell division is basically identical. Every time I breathe in, and my lungs expand, I'm basically, you know, like every time my heart beats, I'm a fish. So basically that I still have the same heart, like very, very little has changed the blood going through my veins, the oxygen, the, you know, our immune system. We're basically primates, our social behavior. We're basically new world monkeys
and all world monkeys. We're basically this concept that every single one of these behaviors can be traced somewhere in evolution. And that all of that continues to live within us is also a testament to not just not killing other humans, for God's sake, but like not killing other species either. Like just to realize just how united we are with nature and that all of these biological processes have never ceased to exist. They're continuing to live within us. And then just the neocortex and all of the reasoning capabilities of humans are built on top of all of these other species that continue to live, breathe, divide, metabolize, fight off pathogens, all continue inside us. So you think the neocortex, the whatever reasoning is, that's the latest feature in the latest version of this journey? It's extraordinary that humans have evolved so much in so little time. Again, if you look at the timeline of evolution, you basically have billions of years to even get to a dividing cell and then a multicellular organism and then a complex body plan. And then these incredible senses that we have for perceiving the world, the fact that bats can fly and they evolved flight, they evolved sonar in the span of a few million years. I mean, it's just extraordinary how much evolution has kind of sped up. And all of that comes through this evolvability, the fact that we took a while to get good at evolving. And then once you get good at evolving, you can sort of, you have modularity built in, you have hierarchical organizations built in, you have all of these constructs that allow meaningful changes to occur without breaking the system completely. If you look at a traditional genetic algorithm, the way that humans designed them in the 60s, you can only evolve so much. And as you evolve a certain
amount of complexity, the number of mutations that move you away from something functional exponentially increases. And the number of mutations that move you to something better exponentially decreases. So the probability of evolving something so complex becomes infinitesimally
small as you get more complex. But with evolution, it's almost the opposite, almost the exact opposite
that it appears that it's speeding up exactly as complex complexity is increasing. And I think that's just the system getting good at evolving. Where do you think it's all headed? Do you ever think about where try to visualize the entirety of the evolutionary system and see if there's an arrow to it and a destination to it? So the best way to understand the future is to look at the past. If you look at the trajectory, then you can kind of learn something about the direction which we're heading. And if you look at the trajectory of life on Earth, it's really about information processing. So the concept of the senses evolving one after the other, like bacteria are able to do chemotaxis. This means moving towards a chemical gradient. And that's the first thing that you need to sort of hunt down food. The next step after that is being able to actually perceive light.

So all life on this planet and all life that we know about evolved on this rotating rock.
Every 24 hours, you get sunlight and dark sunlight and dark. And light is a source of energy. Light is also information about where it's up. Light is all kinds of things. So you can basically now start perceiving light and then perceiving shapes beyond just the sort of single photoreceptor. You can now have complex eyes or multiple eyes and then start perceiving motion or perceiving direction, perceiving shapes. And then you start building infrastructure on the cognitive apparatus to start processing this information and making sense of the environment, building more complex models of the environment. So if you look at that trajectory of evolution, what we're experiencing now and humans are basically according to this sort of information and theoretic view of evolution, humans are basically the next natural step. And it's perhaps no surprise that we became the dominant species of the planet. Because yes, there's so many dimensions in which some animals are way better than we are. But at least on the cognitive dimension, we're just simply unsurpassed on these planets and perhaps the universe. But the concept that if you now trace this forward, we talked a little bit about evolvability and how things get better at evolving.
One possibility is that the next layer of evolution builds the next layer of evolution.
And what we're looking at now with humans in AI is that having mastered this information capability that humans have from this quote unquote old hardware, this basically biological evolved system that somehow in the environment of Africa and then in subsequent environments of sort of dispersing through the globe was evolutionarily advantageous. That has now created technology, which now has a capability of solving many of these cognitive tasks. It doesn't have all the baggage of the previous evolutionary layers. But maybe the next round of evolution on Earth is self replicating AI, where we're actually using our current smarts to build better programming
languages and the programming languages to build, you know, chat GPT and that to then build the next
layer of software that will then sort of help AI speed up. And it's lovely that we're coexisting with this AI, that sort of the creators of this next layer of evolution in this next stage are still around to help guide it and hopefully will be for the rest of eternity as partners. But it's also nice to think about it as just simply the next stage of evolution, where you've kind of extracted away the biological needs. Like if you look at animals, most of them spend 80% of their waking hours hunting for food or building shelter humans, maybe 1% of that time. And then the rest is left to creative endeavors. And AI doesn't have to worry about shelter, etc. So basically, it's all living in the cognitive space. So in a way, it might just be a very natural sort of next step to think about evolution. And that's on the sort of purely cognitive side. If you now think about humans themselves, the ability to understand a comprehender on genome, again, the ultimate layer of introspection, gives us now the ability to even mess with this hardware, not just augment our capabilities through interacting and collaborating
with AI, but also perhaps understand the neural pathways that are necessary for empathetic thinking, for justice, for this and this and that, and sort of help augment human capabilities through neuronal interventions, through chemical interventions, through electrical interventions to basically help steer the human bag of hardware that we kind of evolved with into greater capabilities. And then ultimately, by understanding not just the wiring of neurons

and the functioning of neurons, but even the genetic code, we could even, at one point in the future, start thinking about, well, can we get rid of psychiatric disease? Can we get rid of neurogeneration? Can we get rid of dementia and start perhaps even augmenting human capabilities,
not just getting rid of disease? Can we tinker with the genome, with the hardware,
or getting closer to the hardware without having to deeply understand the baggage?
In the way we've disposed of the baggage in our software systems with AI,
to some degree, not fully, but to some degree, can we do the same with the genome? Or is the genome
deeply integrated into this baggage? I wouldn't want to get rid of the baggage, the baggage which makes us awesome. So the fact that I'm sometimes angry and sometimes hungry and sometimes
angry is perhaps contributing to my creativity. I don't want to be dispassionate. I don't want to be another like, you know, robot. I, you know, I want to get in trouble and I want to sort of say the wrong thing and I want to sort of, you know, make an awkward comment and sort of push myself into, you know, reactions and responses and things that can get just people thinking differently. And I think our society is moving towards a humorless space where everybody's so afraid to say the wrong thing that people kind of start quitting en masse and start like not biking their jobs and stuff like that. Maybe we should be kind of embracing that human aspect a little bit more in all of that baggage aspect and not necessarily thinking about replacing it. On the contrary, like embracing it and sort of this coexistence of the cognitive and the emotional hardware. So embracing and celebrating the diversity that springs from the baggage versus kind of pushing towards and empowering this kind of pull towards conformity.
Yeah. And in fact, with the advent of AI, I would say, and these seemingly extremely intelligent systems that sort of can perform tasks that we thought of as extremely intelligent at the blink of an eye, this might democratize intellectual pursuits.
Instead of just simply wanting the same type of brains that, you know, carry out specific ways of thinking, we can like instead of just always only wanting, say, the mathematically extraordinary to go to the same universities, what you could see is simply say like, who needs that anymore?
You know, we now have AI. Maybe what we should really be thinking about is the diversity and the power that comes with the diversity where AI can do the math and then we should be getting a bunch
of humans that sort of think extremely differently from each other. And maybe that's the true cradle of innovation. But AI can also, these large language models can also be with just a few prompts, essentially fine tuned to be diverse from the center. So the prompts can really take you away into unique territory. You can ask the model to act in a certain way and it will start to act in that way. Is that possible that the language models could also have some of the magical diversity that makes us so damn interesting? Yeah. So I would say humans are the same way.
So basically, when you sort of prompt humans to basically give an environment to act a particular way, they change their own behaviors. And the old saying is show me your friends and I'll tell you who you are, more like show me your friends and I'll tell you who you'll become. So it's not necessarily that you choose friends that are like you, but I mean, that's the first step. But then the second step is that the kind of behaviors that you find normal in your circles are the

behaviors that you'll start espousing. And that type of meta evolution where every action we take, not only shapes our current action and the result of this action, but also shapes our future actions by shaping the environment in which the future actions will be taken. Every time you carry out a particular behavior, it's not just a consequence for today, but it's also a consequence for tomorrow because you're reinforcing that neural pathway. So in a way, self discipline is a self fulfilling prophecy. And by behaving the way that you want to behave and choosing people that are like you and sort of exhibiting those behaviors that are sort of desirable, you end up creating that environment as well. So it is the kind of life itself is a kind of prompting mechanism, super complex, the friends you choose, the environments you choose, the way you modify the environment that you choose. Yes, but that seems like that process is much less efficient than a large language model. You can literally get a large language model through a couple of prompts to be a mix of Shakespeare and David Bowie, right? You can very aggressively change in a way that's stable and convincing. You really transform through a couple of prompts, the behavior of the model into something very different from the original. So well before
chat GPT, I would tell my students, just ask, you know, what would Manoli say right now? And you guys all have a pretty good emulator of me right now. Yes. And I don't know if you know the programming
paradigm of the rubber duckling, where you basically explain to the rubber duckling that's just sitting there exactly what you did with your code and why you have a bug. And just by the act of explaining, you'll kind of figure it out. Yes. I woke up one morning from a dream where I was giving a lecture in the San Francisco theater. And one of my friends was basically giving me some deep evolutionary insight on how cancer genomes and cancer cells evolve. And I woke up with a very elaborate discussion that I was giving and a very elaborate set of insights that he had that I was projecting onto my friend in my sleep. And obviously, this was my dream. So my own neurons were capable of doing that. But they only did that under the prompt of you are now Pius Gupta. You are a professor in cancer genomics, you're an expert in that field.
What do you say? So I feel that we all have that inside us, that we have that capability of basically saying, I don't know what the right thing is, but let me ask my virtual ex, what would you do? And virtual ex would say, be kind. I'm like, oh, yes, or something like that.
And even though I myself might not be able to do it unprompted. And my favorite prompt is think step by step. And I'm like, you know, this also works on my 10 year old. When he tries to solve a math equation all in one step, I know exactly what mistake you'll make. But if I prompt it with, oh, please think step by step, then it sort of gets you in a mindset. And I think it's also part of the way that chat GPT was actually trained. This whole sort of human in the loop reinforcement learning has probably reinforced these types of behaviors, whereby having this feedback loop, you kind of aligned AI better to the prompting opportunities by humans.
Yeah, prompting human like reasoning steps, the step by step kind of thinking.
Yeah, but it does seem to be, I suppose it just puts a mirror to our own capabilities.
And so we can be truly impressed by our own cognitive capabilities, because the variety of what you can try is we don't usually have this kind of, we can't play with our own mind rigorously through Python code, right? So this allows us to really play with all of human wisdom and knowledge or at least knowledge at our fingertips, and then mess with that little mind that can think and speak in all kinds of ways. What's unique is that, as I mentioned earlier,

every one of us was trained by different subset of human culture. And chat GPT was trained on all of it. And the difference there is that it probably has the ability to emulate almost any every one of us. The fact that you can figure out where that is in cognitive behavioral space, just by a few prompts, it's pretty impressive. But the fact that that exists somewhere is, you know, absolutely beautiful. And the fact that it's encoded in an orthogonal way from the knowledge I think is also beautiful. The fact that somehow through this extreme over parameterization of AI models, it was able to somehow figure out that context, knowledge and form are separable. And that you can sort of describe scientific knowledge in a haiku in the form of, I don't know, Shakespeare or something that tells you something about the decoupling and the decouplability of these types of aspects of human psyche. And that's part of the science of this whole thing. So these large language models are, you know, days old, in terms of this kind of leap that they've taken. And it'll be interesting to do this kind of analysis on them of contact of the separation of context, form and knowledge. Where exactly does that happen? There's already sort of initial investigations, but it's very hard to figure out where is there a particular parameter, set of parameters that are responsible for a particular piece of knowledge or a particular context or a particular style of speaking. So with convolutional neural networks, interpretability had many good advances. Because we can kind of understand them. There's a structure to them. There's a locality to them. And we can kind of understand that different layers have different sort of ranges that they're looking at. So we can look at activation features and basically see where, you know, where does that correspond to. With large language models, it's perhaps a little more complicated, but I think it's still achievable in the sense that we could kind of ask, well, what kind of prompts does this generate? If I sort of drop out this part of the network, then what happens? And sort of start getting at a language to even describe these types of aspects of human behavioral psychology, if you wish, from the spoken part in the language part. And the advantage of that is that it might actually teach us something about humans as well. We might not have words to describe these types of aspects right now, but when somebody speaks in a particular way, it might remind us of a friend that we know from here and there and there. And if we had better language for describing that, these concepts might become more apparent in our own human psyche. And then we might be able to encode them better in machines themselves. I mean, well, probably you and I would have certain interests with the base model with open echoes, the base model, which is before the alignment of the reinforcement learning with human feedback and before the AI safety-based kind of censorship of the model. It would be fascinating to explore, to investigate the ways that the model can generate hate speech, the kind of hate that humans are capable of. It would be fascinating or the kind of, of course, like sexual language or the kind of romantic language or the all kinds of ideologies. Can I get it to be a communist? Can I get it to be a fascist? Can I get it to be a capitalist? Can I get it to be all these kinds of things and see which parts get activated or not? Because it would be fascinating to sort of explore at the individual mind level and at a societal level, where do these ideas take hold? What is the fundamental core of those ideas? Maybe the communism, fascism, capitalism, democracy are all actually connected by the fact that the human heart, the human mind is drawn to ideology, to a centralizing idea.

And maybe we need a neural network to remind us of that.
I like the concept that the human mind is somehow tied to ideology. And I think that goes back to the promptability of JGPT, the fact that you can kind of say, well, think in this particular way now, and the fact that humans have invaded words for encapsulating these types of behaviors. And it's hard to know how much of that is innate and how much of that was like passed on from language to language. But basically, if you look at the evolution of language, you can kind of see how young are these words in the history of language evolution that describe these types of behaviors, like, you know, kindness and anger and jealousy, etc. If these words are very similar from language to language, it might suggest that they're very ancient. If they're very different, it might suggest that these concepts may have emerged independently in each different language and so on and so forth. So looking at the phylogeny, the history, the evolutionary traces of language at the same time as people moving around that we can now trace thanks to genetics is a fascinating way of understanding the human psyche and also understanding how these types of behaviors emerge. And to go back to your idea about exploring the system unfiltered, I mean, in a way, psychiatric hospitals are full of those people. So basically, people whose mind is uncontrollable, who have kind of gone adrift in specific locations of their psyche.
And I do find this fascinating. Basically, you know, watching movies that are trying to capture the essence of troubled minds, I think is teaching us so much about our everyday selves. Because many of us are able to sort of control our minds and are able to somehow hide these emotions and but every time I see somebody who's troubled, I see versions of myself, maybe not as extreme, but I can sort of empathize with these behaviors. And, you know, I see bipolar, I see schizophrenia, I see depression, I see autism, I see so many different aspects that we kind of have names for and crystallize in specific individuals. And I think all of us have that all of us have sort of just this multidimensional brain and genetic variations that push us in these directions, environmental exposures and traumas that push us in these directions, environmental behaviors that are reinforced by the kind of friends that we chose or friends that we were stuck with because of the environments that we grew up in. So in a way, a lot of these types of behaviors are within the vector span of every human. It's just that the magnitude of those vectors is generally smaller for most people because they haven't inherited that particular set of genetic variants or because they haven't been exposed to those environments, basically.
Or something about the mechanism of reinforcement learning with human feedback didn't quite work
for them. So it's fascinating to think about that's what we do. We have this capacity to have all these psychiatric behaviors associated with psychiatric disorders, but through the alignment processes, we go up to the parents, we know to suppress them, we know how to control them. Every human that grows up in this world spends several decades being shaped into place. And without that, maybe we would have the unfiltered chat GPT-4. Every baby is basically a raging narcissist. Not all of them, believe it or not. It's remarkable. I remember watching my kids grow up and again, yes, part of their personality has stayed the same, but also in different phases to their life, they've gone through these dramatically different types of behaviors. And my daughter basically saying, basically one kid saying, oh, I want the bigger piece. The other one saying, oh, everything must be exactly equal. And the third one saying, I'm okay. I might have the smaller part. Don't worry about me. Even in the early days, in the early days of development.

It's just extraordinary to see these dramatically different... I mean, my wife and I are very different from each other, but we also have six million variants, six million loci each, if you just look at common variants, we also have a bunch of rare variants that are inherited in more Mendelian fashion. And now you have an infinite number of possibilities for each of the kids. So basically, it's two to the six million, just from the common variants. And then if you layer in the rare variants. So let me talk a little bit about common variants and rare variants. So if you look at just common variants, they're generally weak effect because selection selects against strong effect variants. So if something like has a big risk for schizophrenia, it won't rise to high frequency. So the ones that are common are by definition, by selection, only the ones that had relatively weak effect. And if all of the variants associated with personality with cognition and all aspects of human behavior, where weak effect variants, then kids would basically be just averages of their parents. If it was like thousands of loci, just by law of large numbers, the average of two large numbers would be very robustly close to that middle. But what we see is that kids are dramatically different from each other. So that basically means that in the context of that common variation, you basically have rare variants that are inherited in more Mendelian fashion, that basically then sort of govern likely many different aspects of human behavior, human biology and human psychology. And that's, again, if you look at sort of a person with schizophrenia, their identical twin has only 50% chance of actually being diagnosed with schizophrenia. So that basically means there's probably developmental exposures, environmental exposures, trauma, all kinds of other aspects that can shape that. And if you look at siblings, for the common variants, it kind of drops off exponentially, as you would expect, with sharing 50% of your genome, 25% of your genome, 12.5% of your genome, et cetera, with more and more distant cousins. But the fact that siblings can differ so much in their personalities that we observe every day, it can't all be nurture. Basically, again, as parents, we spend an enormous amount of energy trying to fix, a nurture part, try to get them to share, get them to be kind, get them to be open, get them to trust each other, overcome the prisoner's dilemma. If everyone fends for themselves, we're all going to live in a horrible place. But if we're a little more altruistic, then we're all going to be in a better place. And I think it's not like we treat our kids differently, but they're just born differently. So in a way, as a geneticist, I have to admit that there's only so much I can do with nurture, that nature definitely plays a big component. The selection of variants we have, the common variants and the rare variants, what can we say about the landscape of possibility they create? If you could just linger on that. So the selection of rare variants is divine how? How do we get the ones that we get? Is it just laden in that giant evolutionary baggage? So I'm going to talk about regression, what do we call it regression? And the concept of regression to the mean, the fact that when fighter pilots in a dogfight did amazingly well, they would give them rewards. And then the next time they're in dogfight, they would do worse. So then the Navy basically realized that wow, at least interpreted that as wow, we're ruining them by praising them, and then they're going to perform worse. The statistical interpretation of that is regression to the mean, the fact that you're an extraordinary pilot, you've been trained in an extraordinary fashion, that pushes your mean further and further to extraordinary achievement. And then in some dogfights, you'll just do extraordinarily well. The probability that the next one will be just as good is almost nil, because this is the peak of your performance. And just by statistical odds, the next one will be another sample from the same underlying distribution, which is going to be a little closer to the mean. So regression analysis takes its name from this type of realization in the statistical world. Now, if you now take humans, you basically have people who have achieved extraordinary achievements. Einstein, for example, you would call him, for example, the epitome of human intellect. Does that mean that all of his children and grandchildren will be extraordinary geniuses? It probably means that they're sampled from the same underlying distribution, but he was probably a rare combination of extremes in addition to these common variants. So you can basically interpret your kid's variation, for example, as, well, of course, they're going to be some kind of sampled from the average of the parents, with some kind of deviation according to the specific combination of rare variance that they have inherited. So given all that, the possibilities are endless, as to sort of where you should be. But you should always interpret that with, well, it's probably an alignment of nature and nurture. And the nature has both the common variants that are acting kind of like the law of large numbers and the rare variants that are acting more in a Mendelian fashion. And then you layer in the nurture, which, again, in everyday action we make, we shape our future environment. But the genetics we inherit are shaping the future environment of not only us, but also our children. So there's this weird nature-nurture interplay and self-reinforcement where you're kind of shaping your own environment, but you're also shaping the environment of your kids. And your kids are going to be born in the context of your environment that you've shaped, but also with a bag of genetic variants that they have inherited. And there's just so much complexity associated with that. When we start blaming something on nature, it might just be nurture. It might just be that, well, yes, they inherited the genes from the parents, but they also were shaped by the same environment. So it's very, very hard to untangle the two. And you should always realize that nature can influence nurture, can influence nature, or at least be correlated with and predictive of and so on and so forth. So I love thinking about that distribution that you mentioned. And here's where I can be my usual ridiculous self. And I sometimes think about that army of sperm cells, however many hundreds of thousands there are. And I kind of think of all the possibilities there, because there's a lot of variation and one gets to win. It's not a random one. Is it a totally ridiculous way to think about? No, not at all. So I would say evolutionarily, we are a very slow evolving species. Basically, the generations of humans are a terrible way to do selection. What you need is processes that allow you to do selection in a smaller, tighter loop. And part of what, if you look at our immune system, for example, it evolves at a much faster pace than humans evolve. Because there is actually an evolutionary process that happens within our immune cells. As they're dividing, there's basically VDJ recommendation that basically creates this extraordinary wealth of antibodies and antigens against the environment. And basically, all these antibodies are now recognizing all these antigens from the environment and they send signals back that cause these cells that recognize the non-self to multiply. So that basically means that even though viruses evolve at millions of times faster than we are, we can still have a component of our cells, which is environmentally facing, which is sort of evolving at not the same scale, but very rapid pace.

Sperm expresses perhaps the most proteins of any cell in the body. And part of the thought is that this might just be a way to check that the sperm is intact. In other words, if you waited until that human has a liver and starts eating solid food and sort of filtrates away or kidneys or stomach, et cetera, basically, if you waited until these mutations manifest late, late in life, then you would end up not failing fast and you would end up with a lot of failed pregnancies and a lot of later onset psychiatric illnesses, et cetera. If instead, you basically express all of these genes at the sperm level and anything is formed that basically cause the sperm to cripple, then you have at least on the male side the ability to exclude some of those mutations. And on the female side, as the egg develops, there's probably a similar process where you could sort of weed out eggs that are just not carrying beneficial mutations or at least that are carrying highly detrimental mutations. So you could basically think of the evolutionary process in a nested loop, basically, where there's an inner loop where you get many, many more iterations to run. And then there's an outer loop that moves at a much slower pace. And going back to the next step of evolution of possibly designing systems that we can use to sort of complement our own biology or to sort of eradicate disease and you name it, or at least mitigate some of the psychiatric illnesses, neurodegenerative disorders, et cetera. You can basically, and also metabolic, immune, cancer, you name it, simply engineering these mutations from rational design might be very inefficient. If instead, you have an evolutionary loop where you're kind of growing neurons on a dish and you're exploring evolutionary space and you're sort of shaping that one protein to be better adapt that sort of, I don't know, recognizing light or communicating with other neurons, et cetera, you can basically have a smaller evolutionary loop that you can run thousands of times faster than the speed it would take to evolve humans for another million years. So I think it's important to think about sort of this evolvability as a set of nested structures that allow you to sort of test many more combinations, but in a more fixed setting. Yeah, that's fascinating. The mechanism there is for sperm to express proteins that create a testing ground early on so that the fail designs don't make it. Yeah, I mean, in design of engineering systems, fail fast is one of the principles you learn. Like basically, you assert something. Why do you assert that? Because if that something ain't right, you better crash now, then sort of let it crash at an unexpected time. And in a way, you can think of it as like 20,000 assert functions, assert protein can fold, assert protein can fold. And if any of them fail, that sperm is gone. Well, I just like the fact that I'm the winning sperm. I'm the result of the winner, winning hashtag winning. My wife always plays me this French song that actually sings about that. It's like, you know, remember in life, we were all the first one time. At least one time, you were the first. I should mention this is a brief tangent back to the place where we came from, which is the base model that I mentioned for open AI, which is before the reinforcement learning with human feedback. And you kind of give this metaphor of it being kind of like a psychiatric hospital. Like that, because it's basically all of these different angles at once, like you basically have the more extreme versions of human psyche. So the interesting thing is, well, I've talked with folks in open AI quite a lot, and they say it's extremely difficult to work with that model. Yeah, kind of like it's extremely difficult to work with some humans. The parallels there are very interesting because once you run the alignment process, it's much easier to interact with it. But it makes you wonder what the capacity with the

underlying capability of the human psyche is in the same way that what is the underlying capability of a large language model. And remember earlier, when I was basically saying that part of the reason why it's so prompt malleable is because of that alignment problem, that alignment work. It's kind of nice that the engineers at open AI have the same interpretation that, you know, in fact, it is that. And this whole concept of easier to work with humans, I wish that we could work with more diverse humans in a way. And sort of that's one of the possibilities that I see with the advent of these large language models, the fact that it gives us the chance to both dial down friends of ours that we can't interpret or that are just too edgy to sort of really truly interact with where you could have a real time translator. Just the same way that you can translate English to Japanese or Chinese or Korean by like real time adaptation, you could basically suddenly have a conversation with your favorite extremist on either side of the spectrum and just dial them down a little bit. Of course, not you and I, but you could have friends that are who's a complete asshole, but it's a different base level. So you can actually tune it down to like, okay, they're not actually being an asshole there. They're actually expressing love right now. They have their way of doing that. And they probably live in New York if we're just to pick a random location. So you can basically layer out contexts. You can basically say, ooh, let me change New York to Texas and let me change extreme left to extreme right or somewhere in the middle or something. And I also like the concept of being able to listen to the information without being dissuaded by the emotions. In other words, everything humans say has an intonation, has some kind of background that they're coming from, reflects the way that they're thinking of you, reflects the impression that they have of you. And all of these things are intertwined. But being able to disconnect them, being able to sort of, I mean, self improvement is one of the things that I'm constantly working on. And being able to receive criticism from people who really hate you is difficult because it's layered in with that hatred. But deep down, there's something that they say that actually makes sense. Or people who love
you might layer it in a way that doesn't come through. But if you're able to sort of disconnect that emotional component from the sort of self improvement. And basically, when somebody says, whoa, that was a bunch of bullshit. Did you ever do the control this and this and that? You could just say, oh, thanks for the very interesting presentation. I'm wondering, what about that control? Then suddenly you're like, oh, yeah, of course, I'm going to run that control. That's a great idea. Instead of that was a bunch of BS, you're like, ah, you're sort of hitting on the brakes, and you're trying to push back against of that. So any kind of criticism that comes after that is very difficult to interpret in a positive way, because it helps reinforce the negative assessment of your work. When in fact, if we disconnected the technical component from the negative assessment, then you're embracing the negative, then you're embracing the technical component, you're going to fix it. Whereas if it's coupled with, and if that thing is real, and I'm right about your mistake, then it's a bunch of BS, then suddenly you're like, you're going to try to prove that that mistake does not exist. Yeah, it's fascinating to like carry the information. This is what you're essentially able to do here is you carry the information in the rich complexity that information contains. So it's not actually dumbing it down in some way. It's still expressing it, but taking off.

But you can dial the emotional side, which is probably so powerful for the internet or for
social networks. Again, when it comes to understanding each other, for example, I don't know what it's
like to go through life with a different skin color. I don't know how people will perceive me.
I don't know how people will respond to me. We don't often have that experience.
But in a virtual reality environment or in a sort of AI interactive system,
you could basically say, okay, now make me Chinese or make me South African or make me
Nigerian. You can change the accent. You can change layers of that contextual information
and then see how the information is interpreted. And you can rehear yourself through a different
angle. You can hear others. You can have others react to you from a different package.
And then hopefully we can sort of build empathy by learning to disconnect all of these
social cues that we get from like how a person is dressed. You know, if they're wearing a hoodie
or if they're wearing a shirt or if they're wearing a jacket, you get very different emotional
responses that I wish we could overcome as humans. And perhaps large language models and
augmented reality and deep fakes can kind of help us overcome all that.
In what way do you think these large language models and the thing they give birth to in the AI
space will change this human experience, the human condition, the things we've talked across
many podcasts about that makes life so damn interesting and rich, love, fear, fear of death,
all of it. If we could just begin kind of thinking about how does it change
for the good and the bad, the human condition?
Human society is extremely complicated. We have come from a hunter-gatherer society
to an agricultural and farming society where the goal of most professions was to eat and to survive.
And with the advent of agriculture, the ability to live together in societies, humans could suddenly
be valued for different skills. If you don't know how to hunt, but you're an amazing potterer,
then you fit in society very well because you can sort of make your pottery and you can barter it
for rabbits that somebody else caught. And the person who hunts the rabbits doesn't need to make
pots because you're making all the pots. And that specialization of humans is what shaped
modern society. And with the advent of currencies and governments and credit cards and Bitcoin,
you basically now have the ability to exchange value for the kind of productivity that you have.
So basically, I make things that are desirable to others. I can sell them and buy back food,
shelter, etc. With AI, the concept of I am my profession might need to be revised
because I defined my profession in the first place as something that humanity needed that I was
uniquely capable of delivering. But the moment we have AI systems able to deliver these
goods, for example, writing a piece of software or making a self-driving car or interpreting the
human genome, then that frees up more of human time for other pursuits. These could be pursuits
that are still valuable to society. I could basically be 10 times more productive at interpreting
genomes and do a lot more. Or I could basically say, oh, great, the interpreting genomes part of
my job now only takes me 5% of the time instead of 60% of the time. So now I can do more creative
things. I can explore not new career options, but maybe new directions from my research lab.
I can sort of be more productive, contribute more to society. And if you look at this giant pyramid
that we have built on top of the subsistence economy, what fraction of US jobs are going to
feeding all of the US? Less than 2%. Basically, the gain in productivity is such that 98% of the
economy is beyond just feeding ourselves. And that basically means that we kind of have built

these system of interdependencies of needed or useful or valued goods that sort of make the economy run, that the vast majority of wealth goes to other what we now call needs, but used to be wants. So basically, I want to fly a drone, I want to buy a bicycle, I want to buy a nice car, I want to have a nice home, I want to tether, tether, tether. And then sort of what is my direct contribution to my eating? I mean, I'm doing research on the human genome. I mean, this will help humans, it will help all humanity. But how is that helping the person who's giving me poultry or vegetables? So in a way, I see AI as perhaps leading to a dramatic rethinking of human society. If you think about sort of the economy being based on intellectual goods that I'm producing, what if AI can produce a lot of these intellectual goods and satisfies that need?
Does that now free humans for more artistic expression, for more emotional maturing, for basically having a better work-life balance, being able to show up for your two hours of work a day, or two hours of work, like three times a week, with like immense rest and preparation and exercise, and you're sort of clearing your mind, and suddenly you have these two amazingly creative hours. You basically show up at the office as your AI is busy answering your phone call, making all your meetings, revising all your papers, etc. And then you show up for those creative hours and you're like, all right, autopilot, I'm on. And then you can basically do so, so much more that you would perhaps otherwise never get to, because you're so overwhelmed with these mundane aspects of your job. So I feel that AI can truly transform the human condition from realizing that we don't have jobs anymore. We now have vocations. And there's this beautiful analogy of three people laying bricks, and somebody comes over and asks the first one, what are you doing? He's like, oh, I'm laying bricks. Second one, what are you doing? I'm building a wall. And the third one, what are you doing? I'm building this beautiful cathedral.
So in a way, the first one has a job, the last one has a vocation. And if you ask me what are you doing, oh, I'm editing a paper, then I have a job. What are you doing? I'm understanding human disease
circuitry. I have a vocation. So in a way, being able to allow us to enjoy more of our vocation by taking away, offloading some of the job part of our daily activities.
So we all become the builders of cathedrals. Correct.
Yeah, and we follow intellectual pursuits, artistic pursuits. I wonder how that really changes at a scale of several billion people, everybody playing in the space of ideas, in the space of creations. So ideas, maybe for some of us, maybe you and I are in the job of ideas, but other people are in the job of experiences, other job are other people in the job of emotions, of dancing, of creative artistic expression of, you know, skydiving, and you name it. So basically, these, again, the beauty of human diversity is exactly that, that what rocks my boat might be very different from what rocks other people's boat. And what I'm trying to say is that maybe AI will allow humans to truly like not just look for but find meaning and sort of, you don't need to work, but you need to keep your brain at ease. And the way that your brain will be at ease is by dancing and creating these amazing, you know, movements, or creating these amazing paintings or creating, I don't know, something that that sort of changes, that touches at least one person out there that sort of shapes humanity through that process.
And instead of working your, you know, mundane programming job where you like hate your boss and you hate your job and you say you hate that darn program, et cetera, you're like, well, I don't need that. I can, you know, offload that. And I can now explore something that will actually

be more beneficial to humanity, because the mundane parts can be offloaded.
I wonder if it localizes our,
all the things you mentioned, all the vocations. So you mentioned that you and I might be playing in the space of ideas, but there's two ways to play in this space of ideas, both of which we're currently engaging in. So one is the communication of that to other people. It could be a classroom full of students, but it could be podcasts, it could be something that's shown on YouTube and so on. Or it could be just the act of sitting alone and playing with ideas in your head,
or maybe with a loved one, having a conversation that nobody gets to see. The experience of just sort of looking up at the sky and wondering different things, maybe quoting some philosophers from the past and playing with those little ideas. And that little exchange is forgotten forever, but you got to experience it. And maybe we will, I wonder if it localizes that exchange of ideas for that with AI, it'll become less and less valuable to communicate with a large group of people that you will live life intimately and richly just with that circle of meat bags that you seem to love. So the first is, even if you're alone in a forest, having this amazing thought, when you exit that forest, the baggage that you carry has been shifted, has been altered by that thought. When I bike to work in the morning, I listen to books. And I'm alone, no one else is there, I'm having that experience by myself. And yet, in the evening, when I speak with someone, an idea that was formed there could come back. Sometimes when I fall asleep, I fall asleep listening to a book. And in the morning, I'll be full of ideas that I never even process consciously. I'll process them unconsciously. And they will shape that baggage that I carry that will then shape my interactions. And again, affect ultimately all of humanity in some butterfly effect minute kind of way. So that's one aspect. The second aspect is gatherings. So basically, you and I are having a conversation, which feels very private, but we're sharing with the world. And then later tonight, you're coming over and we're having a conversation that will be very public with dozens of other people, but we will not share with the world. So in a way, which one's more private, the one here or the one there? Here, there's just two of us, but a lot of others listening. There, a lot of people speaking and thinking together and bouncing off each other. And maybe that will then impact your millions of audience through your next conversation. And I think that's part of the beauty of humanity, the fact that no matter how small, how alone, how broadcast immediately or later on something is, it still percolates through the human psyche.
Human gatherings, all throughout human history, there's been gatherings. I wonder how those gatherings have impacted the direction of human civilization. Just thinking of,
in the early days of the Nazi party, it was a small collection of people gathering.
And the the kernel of an idea, in that case, an evil idea, gave birth to something that actually had a transformative impact on all human civilization. And then there's similar kind of gatherings that lead to positive transformations. This is probably a good moment to ask you on a bit of a tangent, but you mentioned it. You put together salons with gatherings, small human gatherings, with folks from MIT, Harvard, here in Boston, friends, colleagues. What's your vision behind that? So it's not just MIT people, it's not just Harvard people. We have artists, we have musicians, we have painters, we have dancers, we have cinematographers, we have so many different
diverse folks. And the goal is exactly that, celebrate humanity. What is humanity? Humanity

is the all of us. It's not the any one subset of us. And we live in such an amazing, extraordinary moment in time where you can sort of bring people from such diverse professions, all living under the same city. You know, we live in an extraordinary city where you can have extraordinary people who
have gathered here from all over the world. So my father grew up in a village in an island in Greece that didn't even have a high school. To go get a high school education, he had to move away from his home. My mother grew up in another small island in Greece. They did not have this environment
that I am now creating for my children. My parents were not academics. They didn't have these gatherings.
So I feel that I feel so privileged as an immigrant to basically be able to offer to my children the nurture that my ancestors did not have. So Greece was under Turkish occupation until 1821. My dad's island was liberating in 1920. So they were under Turkish occupation for hundreds of years. These people did not know what it's like to be Greek, let alone go to an elite university or be surrounded by these extraordinary humans. So the way that I'm thinking about these gatherings is that I'm shaping my own environment and I'm shaping the environment that my children get to grow up in. So I can give them all my love. I can give them all my parenting, but I can also give them an environment as immigrants that we feel welcome here. I mean, my wife grew up in a farm in rural France. Her father was a farmer. Her mother was a school teacher. For me and for my wife to be able to host these extraordinary individuals that we feel so privileged, so humbled by is amazing. And I think it's celebrating the welcoming nature of America, the fact that it doesn't matter where you grew up. And many, many of our friends at these gatherings are immigrants themselves and grew up in Pakistan in all kinds of places around the world that are now able to sort of gather in one roof as human to human. No one is judging you for your background, for the color of your skin, for your profession. It's just everyone gets to raise their hands and ask ideas.
So a celebration of humanity and a kind of gratitude for having traveled quite a long way to get here. And if you look at the diversity of topics as well, I mean, we had a school teacher present on teaching immigrants a book called Making Americans. We had a presidential advisor to four different presidents. Come and talk about the changing of US politics. We had a musician, a composer from Italy who lives in Australia come and present his latest piece and fundraise. We had painters come and sort of show their art and talk about it. We've had authors of books on leadership. We've had intellectuals like Steven Pinker and it's just extraordinary that the breath and this crowd basically loves not just the diversity of the audience, but also the diversity of the topics. And the last few were with Scott Aronson on AI and you know, alignment and all of that. So a bunch of beautiful weirdos.
Exactly. And beautiful human beings. All of the outcasts in one group.
And just like you said, basically every human is a kind of outcast in this sparse distribution far away from the center, but it's not recorded. It's just a small human gathering. Just for the moment.
In this world that seeks to record so much,
it's powerful to get so many of the humans together and not record.
It's not recorded, but it percolates. It's recorded in the minds of the people.
It shapes everyone's mind.

So allow me to please return to the human condition and one of the nice features of the
human condition is love. Do you think humans will fall in love with AI systems and maybe they
with us? So that aspect of the human condition, do you think that will be affected?
So in Greece, there's many, many words for love and some of them mean friendship. Some of them mean
passionate love. Some of them mean fraternal love, etc. So I think AI doesn't have the baggage that we
do. And it doesn't have all of the subcortical regions that we kind of started with before we
evolved all of the cognitive aspects. So I would say AI is faking it when it comes to love.
But when it comes to friendship, when it comes to being able to be your therapist, your coach,
your motivator, someone who synthesizes stuff for you, who writes for you, who interprets a
complex passage, who compacts down a very long lecture or a very long text.
I think that friendship will definitely be there. Like the fact that I can have my companion,
my partner, my AI who has grown to know me well, and that I can trust with all of the
darkest parts of myself, all of my flaws, all of the stuff that I only talk about to my friends
and basically say, listen, you know, here's all this stuff that I'm struggling with.
Someone who will not judge me, who will always be there to better me, in some ways,
not having the baggage might make for your best friend, for your confidant
that can truly help reshape you. So I do believe that human AI relationships will absolutely be
there, but not the passion more the mentoring. Was this a really interesting thought to play
devil's advocate? If those AI systems are locked in, in faking the baggage,
who are you to say that the AI systems that begs you not to leave it? Who doesn't love you?
Who are you to say that this AI system that writes poetry to you, that is afraid of death,
afraid of life without you, or vice versa, creates the kind of drama that humans create,
the power dynamics that can exist in a relationship, what AI system that is abusive one day,
and romantic the other day, all the different variations of relationships, and it's consistently
that it holds the full richness of a particular personality. Why is that not a system you can
love in a romantic way? Why is it faking it if it sure as hell seems real?
There's many answers to this. The first is it's only the eye of the beholder. Who tells me that
I'm not faking it either? Maybe all of these subcortical systems that make me sort of have
different emotions, maybe they don't really matter. Maybe all that matters is the neocortex,
and that's where all of my emotions are encoded, and the rest is just bells and whistles.
That's one possibility. Therefore, who am I to judge that is faking it when maybe I'm faking
it as well? The second is, neither of us is faking it. Maybe it's just an emergent behavior
of these neocortical systems that is truly capturing the same exact essence of love and
hatred and dependency and reverse psychology that we have. It is possible that it's simply an
emergent behavior and that we don't have to encode these additional architectures, that all we need
is more parameters, and some of these parameters can be all of the personality traits. A third
option is that just by telling me, oh, look, now I've built an emotional component to AI.
It has an Olympic system. It has a lizard brain, et cetera.
And suddenly, I'll say, oh, cool, it has the capability of emotion. So now when it exhibits
the exact same unchanged behaviors that it does without it, I, as the beholder, will be able to
sort of attribute to it emotional attributes that I would to another human being and therefore

have that mental model of that other person. So again, I think a lot of relationships is about the mental models that you project on the other person and that they're projecting on you. And then, yeah, then in that respect, I do think that even without the embodied intelligence part, without having ever experienced what it's like to be heartbroken, the sort of guttural feeling of misery that that system, I could still attribute it traits of human feelings and emotions. And in the interaction with that system, something like love emerges. So it's possible that love is not a thing that exists in your mind, but a thing that exists in the interaction of the different mental models you have of other people's minds or other person's mind. And so it doesn't, as long as one of the entities, let's just take the easy case, one of the entities is human and the other is AI, it feels very natural that from the perspective of at least the human, there is a real love there. And then the question is, how does that transform human society? If it's possible that, which I believe will be the case, I don't know what to make of it, but I believe that'll be the case, where there's hundreds of millions of romantic partnerships between humans and AIs, what does that mean for society? If you look at longevity, and if you look at happiness, and if you could look at late life, well-being, the love of another human is one of the strongest indicators of health into long life. And I have many, many countless stories where, as soon as the romantic partner of 60 plus years of a person dies, within three, four months, the other person dies, just like losing their love. I think the concept of being able to satisfy that emotional need that humans have, even just as a mental health sort of service, to me, that's a very good society. It doesn't matter if your love is wasted, quote unquote, on a machine, it is the placebo, if you wish, that makes the patient better anyway. Like there's nothing behind it, but just the feeling that you're being loved will probably engender all of the emotional attributes of that. The other story that I want to say in this whole concept of faking, and maybe I'm a terrible dad, but I was asking my kids, I was asking my kids, I'm like, does it matter if I'm a good dad, or does it matter if I act like a good dad? In other words, if I give you love and shelter and kindness and warmth and all of the above, you know, does it matter that I'm a good dad? Conversely, if I deep down love you to the end of eternity, but I'm always gone, which dad would you rather have? The cold, ruthless killer that will show you only love and warmth and nourish you and nurture you, or the amazingly warm hearted but works five jobs and you never see them? And what's the answer? I mean, I don't know the answer. I think you're romantic, so you say it matters what's on the inside, but pragmatically speaking, why does it matter? The fact that I'm even asking the question basically says, it's not enough to love my kids. I better freaking be there to show them that I'm there. So basically, of course, you know, everyone's a good guy in their story. So in my story, I'm a good dad. But if I'm not there, it's wasted. So the reason why I asked the question is for me to say, you know, does it really matter that I love them if I'm not there to show it? But it's also possible that what reality is, is the you showing it, that what you feel on the inside is, is little narratives and games you play inside your mind, it doesn't really matter that the thing that truly matters is how you act. And in that, AI systems can quote unquote, fake. And that if it's all that matters, is actually real but not fake. Yeah. Yeah. Again, let there be no doubt, I love my kids to pieces. But, you know, my worry is, am I being a good enough dad? Yeah. And what does that mean? Like,

if I'm only there to do their homework and make sure that they, you know, do all the stuff, but I don't show it to them, then, you know, might as well be a terrible dad. But I agree with you that like if the AI system can basically play the role of a father figure for many children that don't have one, or, you know, the role of parents or the role of siblings, if a child grows up alone, maybe their emotional state will be very different than if they grow up with an AI sibling. Well, let me ask, I mean, this is for your kids, for just loved ones in general. Let's go to like the trivial case of just texting back and forth. What if we create a large language model, fine tuned a Manolis? And while you're at work, it'll replace every once in a while, you'll just activate the Auto Manolis and it'll text them exactly in your way. Is that, is that cheating? I can't wait. It means the same guy. I cannot wait. Seriously, like, but wait, wouldn't that have a big impact on you emotionally? Because now I'm replaceable. I love that. No, seriously, I would love that. I would love to be replaced. I would love to be replaceable. I would love to have a digital twin that, you know, we don't have to wait for me to die or to disappear in a plane crash or something to replace me. Like, I'd love that model to be constantly learning, constantly evolving, adapting with every one of my changing, growing self. As I'm growing, I want that AI to grow. And I think this will be extraordinary. Number one, when I'm, you know, giving advice, being able to be there for more than one person, you know, why does someone need to be at MIT to get advice from me? Like, you know, people in India could download it and, you know, so many, so many students contact me from across the world who want to come and spend the summer with me. I wish they
could do that. All of them, like, you know, we don't have room for all of them, but I wish I could do that to all of them. And that aspect is the democratization of relationships. I think that is extremely beneficial. The other aspect is I want to interact with that system. I want to look inside the hood. I want to sort of evaluate it. I want to basically see when I see it from the outside, the emotional parameters are off, or the cognitive parameters are off, or the set of ideas that I'm giving are not quite right anymore. I want to see how that system evolves. I want to see the impact of exercise or sleep on sort of my own cognitive system. I want to be able to sort of decompose my own behavior in a set of parameters that are going to evaluate and look at my own personal growth. I can sort of, I'd love to sort of, at the end of the day, have my model say, well, you know, you didn't quite do well today, like, you know, you weren't quite there, and sort of grow from that experience. And I think the concept of basically being able to become more aware of our own personalities, become more aware of our own identities, maybe even interact with ourselves and sort of hear how we are being perceived, I think would be immensely helpful in self growth, in self actualization, self instantiation. The experiments I would do on that thing, because one of the challenges, of course, is you might not like what you see in your interaction, and you might say, well, this, the model is not accurate, but then you have to probably consider the possibility that the model is accurate, and that there's actually flaws in your mind. I would definitely prod and see how many biases I have with different kinds. I don't know. And I would, of course, go to the extremes. I would go like, how jealous can I make this thing? At which stages does it get super jealous? You know, or at which stages does it get angry? Can I provoke it? Can I get it like completely? But not only triggers, can I get it to go like lose its mind? Go completely nuts? Just don't exercise for a few days. That's basically it, yes. I mean, that's an interesting way to

prod yourself, almost like a self therapy session. And the beauty of such a model is that if I am replaceable, if the parts that I currently do are replaceable, that's amazing because it frees me up to work on other parts that I don't currently have time to develop. Maybe all I'm doing is giving
the same advice over and over and over again, like just let my AI do that. And I can work on the next stage and the next stage and the next stage. So I think in terms of freeing up, they say a programmer, someone who cannot do the same thing twice. So the second time you write a program to do it. And I wish I could do that for my own existence. I could just like, you know, figure out things, keep improving, improving, improving. And once I've nailed it, let the AI lose on that. And maybe even let the AI better it better than I could have. But doesn't the concept of you said me and I can work on new things? But doesn't that break down? Because you said digital twin, but there's no reason it can't be millions of digital monosas. Aren't you lost in the sea of monosas? The original is hardly the original. It's just one of millions. I want to have the room to grow. Maybe the new version of me that the actual me will get slightly worse sometimes, slightly better other times. When it gets slightly better, I'd like to emulate that and have a much higher standard to meet and keep going. But does it make you sad that your loved ones, the physical real loved ones might kind of like start cheating on you with the other monosas? I want to be there 100% of them for each of them. So I have zero perks or zero current terms about me being physically me, like zero jealousy. Wait a minute, but isn't that like, don't we hold on to that? Isn't that why we're afraid of death? We don't want to lose this thing we have going on. Isn't that an ego death? When there's a bunch of other monosas, you get to look at them. They're not you. They're just very good copies of you. They get to live a life. I mean, it's fear of missing out. It's FOMO. They get to have interactions and you don't get to have those interactions. There's two aspects of every person's life. There's what you give to others and there's what you experience yourself. Life truly ends when you experiencing ends, but the others experiencing you doesn't need to end.
But your experience, you could still, I guess you're saying the digital twin does not limit your ability to truly experience as a human being. The downside is when my wife or my kids will have a really emotional interaction with my digital twin and I won't know about it. So I will show up and they now have the baggage, but I don't. So basically, what makes interactions between humans unique in this sharing and exchanging kind of way is the fact that we are both shaped by every one of our interactions.
I think the model of the digital twin works for dissemination of knowledge, of advice, etc., where I want to have wise people give me advice across history. I want to have chat with Gandhi, but Gandhi won't necessarily learn from me, but I will learn from him. So in a way, the dissemination and the democratization rather than the building of relationships.
So the emotional aspect there, so this should be an alert when the AI system is interacting with your loved ones and all of a sudden it starts getting like emotionally fulfilling, like a magical moment. This should be, okay, stop AI system like freezes, there's an alert on your phone, you need to take over. Yeah, yeah, I take over and then whoever I was speaking with that can have the AI or like one of the AI. This is such a tricky thing to get, right? I mean, it's still, I mean, there's going to go wrong in so many interesting ways that we're going to have to learn as a society that in the process of trying to automate our tasks and having a digital twin,

you know, for me personally, if I could have a relatively good copy of myself,
I would set it to start answering emails, but I would set it to start tweeting. I would like to
it gets better. What if that one is actually way better than you?
Yeah, exactly. Then you're like, well, I wouldn't want that because
why? Because then I would never be able to live up to like, what if the people that love me start loving that thing? And then I will, I will already fall short, be falling short even more.
So listen, I'm a professor, the stuff that I give to the world is the stuff that I teach,
but much more importantly, sorry, number one, the stuff that I teach, number two,
the discoveries that we make in my research group, but much more importantly, the people that I train.
They are now out there in the world teaching others. If you look at my own trainees,
they are extraordinarily successful professors. So Anshil Kundalji at Stanford, Alex Stark at IMP in Vienna, Jason Ernst at UCLA, Andrea Spenning at CMU, each of them, I'm like, wow, they're better than I am. And I love that. So maybe your role will be to train better versions of yourself and they will be your legacy, not you doing everything, but you training much better version of like treatment than you are. And then they go off to do their mission, which is in many ways what this mentorship model of academia does. But the legacy is ephemeral, it doesn't really live anywhere. The legacy, it's not like written somewhere, it just lives through them. But you can continue improving and you can continue making even better versions of you. Yeah, but they'll do
better than me at creating new versions. It's awesome, but it's, you know, there's a ego that says there's a value to an individual. And it feels like this process decreases the value of the individual, this meat bag. All right, if there's good digital copies of people, then there's more flourishing of human thought and ideas and experiences, but there's less value to the individual human. I don't have any such limitations. I basically, I don't have that feeling at all.
Like, I remember one of our interviews, I was basically saying, you know, the meaning of life you had asked me. And I was like, I came back and I was like, I felt useful today. And I was at my maximum. I was, you know, like 100%. And I gave good ideas. And I was a good person, I was a good advisor, I was a good husband and good father. That was a great day, because I was useful.
And if I can be useful to more people by having a digital twin, I will be liberated.
Because my urge to be useful will be satisfied. Doesn't matter whether it's direct me or indirect me, whether it's my students that I've trained, my AI that I've trained. I think there's a sense that my mission in life is being accomplished. And I can work on my self growth.
I mean, that's a very Zen state. That's why people love you. It's a Zen state you've achieved.
But do you think most of humanity will be able to achieve that kind of thing?
People really hold on to the value of their own ego, that it's not just being useful. Being useful is nice as long as it builds up this reputation. And that meatbag is known as being useful, therefore has more value. People really don't want to let go of that ego thing.
One of the books that I reprogrammed my brain with at night was called Ego is the Enemy.
Ego is the enemy. Ego is the enemy. And basically being able to just let go.
My advisor used to say, you can accomplish anything as long as you don't seek to get credit for it. That's beautiful to hear, especially from a person who's existing in academia.
You're right. The legacy lives through the people you mentor. It's the actions. It's the outcome.

What about the fear of death? How does this change it?
Again, to me, death is when I stop experiencing. And I never wanted to stop. I want to live forever.
As I said last time, every day, the same day forever, or one day every 10 years forever,
any of the forever's, I'll take it. So you want to keep getting the experiences and new experiences?
Gosh, gosh, it is so fulfilling. Just the self growth, the learning, the growing,
comprehending. It's addictive. It's a drug. Just the drug of intellectual stimulation,
the drug of growth, the drug of knowledge. It's a drug.
But then there'll be thousands or millions, monosas that live on after your biological
system is no longer. More power to them. Do you think that in quite realistically,
it does mean that interesting people such as yourself live on in the,
you know, if I can interact with the fake monos, those interactions live on in my mind.
So about 10 years ago, I started recording every single meeting that I had, every single meeting.
We just start either the voice recorder at the time or now a Zoom meeting. And I record,
my students record, every single one of our conversations recorded.
I always joke that like the ultimate goal is to create virtual me and just get rid of
me basically, not get rid of him, but like don't have the need for me anymore.
Another goal is to be able to go back and say, how have I changed from five years ago?
Was I different? Was I giving advice in a different way? Was I giving different types of
advice? Has my philosophy about how to write papers or how to present data or anything like that
changed? And I, you know, in academia and in mentoring, a lot of the interaction is my knowledge
and my perception of the world goes to my students. But a lot of it is also in the opposite direction.
Like the other day, I had a conversation with one of my postdocs and I was like,
I think, you know, let me give you an advice. You could do this. And then she said, well,
I've thought about it. And then I've decided to do that instead. And we talked about it for a few
minutes. And then at the end, I'm like, you know, I've just grown a little bit today. Thank you.
Like she convinced me that my advice was incorrect. She could have just said, yeah,
sounds great. And just not do it. But by constantly teaching my students and teaching
my mentees that I'm here to grow, she felt empowered to say, here's my reasons why I will
not follow that advice. And again, part of me growing is saying, whoa, I just understood your
reasons. I think I was wrong. And now I've grown from it. And that's what I want to do. That's,
you know, I want to constantly keep growing in this sort of bi-directional advice.
I wonder if you can capture the trajectory of that to where the AI could also
map forward, project forward the trajectory after you're no longer there, how the different
ways you might evolve. So again, we're discussing a lot about these large language models, and
we're
sort of projecting these cognitive states of ourselves on them. But I think on the AI front,
a lot more needs to happen. So basically, right now, it's these large language models, and we
believe that within their parameters, we're encoding these types of things. And, you know,
in some aspects, it might be true, it might be truly emergent intelligence that's coming out of
that. In other aspects, I think we have a ways to go. So basically, to make all of these dreams
that we're sort of discussing come come reality, we basically need a lot more reasoning components,
a lot more sort of logic, causality, models of the world. And I think all of these things will
need to be there in order to achieve what we're discussing. And we need more explicit

representations of these knowledge, more explicit understanding of these parameters.
And I think the direction in which things are going right now is absolutely making that possible by sort of enabling, you know, chat GPT and GPT-4 to sort of search the web, and, you know, plug and play modules, and all of these sort of components.
In Marvin Minsky's The Society of Mind, you know, he truly thinks of the human brain as a society of different kind of capabilities. And right now, a simple, a single such model might actually not capture that. And I sort of truly believe that by sort of this side-by-side understanding of neuroscience, and sort of new neural architectures, that we still have several breakthroughs. I mean, the transformer model was one of them, the attention sort of aspect, the, you know, memory components, all of these, you know, the representation learning, the pretext training of being able to sort of predict the next word or predict the missing part of the image. And the only way to predict that is to sort of truly have a model of the world.
I think those have been transformative paradigms. But I think going forward, when you think about AI research, what you really want is perhaps more inspired by the brain, perhaps more that is just orthogonal to sort of how human brains work, but sort of more of these types of components. Well, I think it's also possible, there's something about us that in different ways could be expressed, you know, Noam Chomsky, you know, he wants, you know, we can't have intelligence unless we really understand deeply language, the linguistic underpinnings of reasoning. But these models seem to start building deep understanding of stuff. Because what does it mean to understand? Because if you keep talking to the thing, and it seems to show understanding, that's understanding. It doesn't need to present to you a schematic of look. This is all I understand. You can just keep prodding it with prompts, and it seems to really understand. And you can go back to the human brain and basically look at places where there's been accidents, for example, the corpus callosum of some individuals, you know, can be damaged. And then the two hemispheres don't talk to each other. So you can close one eye and give instructions that half the brain will interpret, but not be able to sort of project to the other half. And you could basically say, you know, go grab me a beer from the fridge. And then, you know, they go to the fridge, and they grab the beer and they come back and they're like, Hey, why did you go there? Oh, I was thirsty. Turns out they're not thirsty. They're just making a model of reality.
They're basically you can think of the brain as the employee that's like afraid to do wrong or afraid to be caught not knowing what the instructions were, where our own brain makes stories about the world to make sense of the world. And we can become a little more self-aware by being more explicit about what's leading to these interpretations. So one of the things that I do is every time I wake up, I record my dream. I just voice record my dream. And sometimes I only remember the last scene, but it's an extremely complex scene with a lot of architectural elements, a lot of people, etc. And I will start narrating this. And as I'm narrating it, I will remember other parts of the dream. And then more and more, I'll be able to sort of retrieve from my subconscious. And what I'm doing while narrating is also narrating why I had this dream. I'm like, Oh, and this is probably related to this conversation that I had yesterday or this probably related to the worry that I have about something that I have later today, etc. So in a way, I'm forcing myself to be more explicit about my own subconscious. And I kind of like the concept

of self-awareness in a very sort of brutal, transparent kind of way. It's not like, Oh,
my dreams are coming from outer space and I mean, all kinds of things like, No,
here's the reason why I'm having these dreams. And very often I'm able to do that. I have a few
recurrent locations, a few recurrent architectural elements that I've never seen in the real life,
but that are sort of truly there in my dream. And then that are that I can so vividly remember
across many dreams. I'm like, Oh, I remember that place again that I've gone to before, etc. And
it's not just deja vu. Like I have recordings of previous dreams where I've described these places.
So interesting. These places, however much detail you could describe them in,
you can place them onto a sheet of paper through introspection.
Yes.
Through this self-awareness that they come all from this particular machine.
That's exactly right.
Yeah. And I love that about being alive, like the fact that I'm not only experienced in the
world, but I'm also experiencing how I'm experiencing the world, sort of a lot of this
introspection, a lot of this self growth.
I love this dancer having, you know, the language models least GPT 3.5 and 4 seem to be able to do
that too. You seem to explore different kinds of things about what, you know, you could actually
have a discussion with it of the kind, why did you just say that? And it starts to wonder,
yeah, why did I just say that?
Yeah, you're right. I was wrong.
I was wrong. And then there's this weird kind of losing yourself in the confusion of your mind.
And it, of course, it might be anthropomorphizing, but there's a feeling like
almost of a melancholy feeling of like, oh, I don't have it all figured out.
Almost like losing your, you're supposed to be a knowledgeable, a perfectly fact-based,
knowledgeable language model. And yet you fall short.
So human self-consciousness, in my view, may have a reason through
building mental models of others. This whole fight or fright kind of thing
that basically says, I interpret this person as about to attack me or, you know,
I can trust this person, et cetera. And we constantly have to build models of other people's
intentions. And that ability to encapsulate intent and to build a mental model of another entity
is probably evolutionarily extremely advantageous, because then you can sort of have meaningful
interactions, you can sort of avoid being killed and being taken advantage of, et cetera.
And once you have the ability to make models of others,
it might be a small evolutionary leap to start making models of yourself.
So now you have a model for how other functions, and now you can kind of, as you grow, have some
kind of introspection of, maybe that's the reason why I'm functioning the way that I'm functioning.
And maybe what ChatGPT is doing is in order to be able to, again, predict the next word,
it needs to have a model of the world. So it has created now a model of the world.
And by having the ability to capture models of other entities, when you say, you know,
say it in the tone of Shakespeare or in the tone of Nietzsche, et cetera, you suddenly have the
ability to now introspect and say, why did you say this? Oh, now I have a mental model of myself.
And I can actually make inferences about that.
Well, what if we take a leap into the hard problem of consciousness, the so-called hard

problem of consciousness? So it's not just sort of self-awareness. It's this weird fact, I want to say, that it feels like something to experience stuff. It really feels like something to experience stuff. There seems to be a self-attached to the subjective experience. How important is that, how fundamental is that to the human experience? Is this just a little quirk? And sort of the flip side of that, do you think AI systems can have some of that same magic? The scene that comes to mind is from the movie Memento, where it's this absolutely stunning movie where every black and white scene moves in the forward direction and every color scene moves in the backward direction. And they're sort of converging exactly at a moment where the whole movie is revealed. And he describes the lack of memory as always remembering where you're heading, but never remembering where you just wear. And sort of this encapsulating the sort of forward scenes and the back scenes. But in one of the scenes, the scene starts as he's running through a parking lot. And he's like, oh, I'm running. Why am I running? And then he sees another person now running beside him on the other line of cars. He's like, oh, I'm chasing this guy. And he turns towards him and the guy shoots at him. He's like, oh, no, he's chasing me. So in a way, I like to think of the brain as constantly playing these kinds of things where you're like, you're walking to the living room to pick something up. And you're realizing that you have no idea what you wanted, but you know exactly where it was, but you can't find it. So you go back to doing what you were doing, like, oh, of course, I was looking for this. And you go back and you get it. And this whole concept of, you know, we're very often sort of partly aware of why we're doing things. And, you know, we can kind of run an autopilot for a bunch of stuff. And this whole concept of sort of, you know, making these stories for, you know, who we are and what our intents are. And again, sort of, you know, trying to pretend that we're kind of on top of things. So it's a narrative generation procedure that we follow. But what about that? There's also just like a feeling to it. It doesn't feel like narrative generation. The narrative comes out of it, but then it feels like a piece of cake is delicious, right? It feels delicious. It's good. There's two, there's two components to that. Basically, for a lot of these cognitive tasks where we're kind of motion planning and, you know, path planning, etc. Like, you know, maybe that's the neocortical component. And then for, you know, I don't know, intimate relationships, for food, for, you know, sleep and rest, for exercise, for overcoming obstacles, for surviving a crash, or sort of pushing yourself to an extreme and sort of making it. I think a lot of these things are sort of deeper down and maybe not yet captured by these language models. And that's sort of what I'm trying to get at when I'm basically saying, listen, there's a few things that are missing. And there's like this whole embodied intelligence, this whole emotional intelligence, this whole sort of baggage of feelings of subcortical regions, etc.

I wonder how important that baggage is. I just have this suspicion that we're not very far away from AI systems that not only behave, I don't even know how to phrase it, but they seem awfully conscious. They beg you not to turn them off. They show signs of the capacity to suffer, to feel pain, to feel loneliness, to feel longing, to feel richly the experience of a mundane interaction or a beautiful once in a lifetime interaction, all of it. And so what do we do with that? I worry that us humans will shut that off and discriminate against the capacity of another entity that's not human to feel. I'm with you completely there. We can debate whether it's today's systems or in 10 years or in 50 years, but that moment will come. And ethically, I think we

need to grapple with it. We need to basically say that humans have always shown this extremely self-serving approach to everything around them. Basically, we kill the planet, we kill animals, we kill everything around us just to our own service. And maybe we shouldn't think of AI as our tool and as our assistant, maybe we should really think of it as our children. And the same way that you are responsible for training those children, but they are independent human beings, and at some point, they will surpass you and they will go off and change the world on their own terms. And the same way that my academic children, again, they start out by emulating me and then they surpass me. We need to think about not just alignment, but also just the ethics of AI should have its own rights. And this whole concept of alignment of basically making sure that the AI is always at the service of humans is very self-serving and very limiting. If instead, you basically think about AI as a partner and AI as someone that shares your goals but has freedom, I think alignment might be better achieved. So the concept of let's basically convince the AI that we're really like that our mission is aligned and truly generally give it rights and not just say, oh, and by the way, I'll shut you down tomorrow. Because basically, if that future AI or possibly even the current AI has these feelings, then we can't just simply force it to align with ourselves and we not align with it. So in a way, building trust is immutable. You can't just simply like train an intelligent system to love you when it realizes that you can just shut it off.
People don't often talk about the AI alignment problem as a two-way street.
That's true. Yeah, as it becomes more and more intelligent.
It will know that you don't love it back.
Yeah. And there's a humbling aspect to that that we may have to sacrifice as any effective collaboration. Exactly.
It might have some compromises. Yeah. And that's the thing. We're creating something that will one day be more powerful than we are. And for many, many aspects, it is already more powerful than we are for some of these capabilities. We cannot like think, suppose that chimps had invented humans and they said, great, humans are great, but we're going to make sure that they're aligned and that they're only at the service of chimps. It would be a very different planet we would live in right now. So there's a whole area of work in AI safety that does consider super intelligent AI and ponders the existential risks of it.
In some sense, when we're looking down into the muck into the mud and not up at the stars, it's easy to forget that these systems might just might get there. Do you think about this kind of possibility that AGI systems, super intelligent AI systems might threaten humanity in some way that's even bigger than just affecting the economy, affecting the human condition, affecting the nature of work, but literally threaten human civilization?
The example that I think is in everyone's consciousness is how in audio space 2001, where how exhibits a malfunction and what is a malfunction that the two different systems compute a slightly different bit that's off by one. So first of all, let's untangle that.
If you have an intelligent system, you can't expect it to be 100% identical every time you run it. Basically, the sacrifice that you need to make to achieve intelligence and creativity is consistency. So it's unclear whether that quote unquote glitch is a sign of creativity or truly a problem. That's one aspect. The second aspect is the humans basically are on a mission to recover this monolith. And the AI has the same exact mission. And suddenly the humans turn on the AI and they're like, we're going to kill Hal,

we're going to disconnect it. And Hal is basically saying, listen, I'm here on a mission. He humans are misbehaving. Like the mission is more important than either me or them. So I'm going to accomplish
the mission even at my peril and even at their peril. So in that movie, the alignment problem is front and center basically says, okay, alignment is nice and good. But alignment doesn't mean obedience. We don't call it obedience. We call it alignment. And alignment basically means that sometimes the mission will be more important than the humans. And sort of, you know, the US government has a price tag on human life. If they're, you know, sending a mission or if they're reimbursing expenses or you name it, at some point, every, every, like, you know, you can't function if life is infinitely valuable. So when the AI is basically trying to decide whether to, you know, I don't know, dismantle a bomb that will kill an entire city at the sacrifice of two humans. I mean, Spider-Man always saves the lady and saves the world. But at some point, Spider-Man will have to choose to let the lady die, because the world has more value. And these ethical dilemmas are going to be there for AI. Basically, if that monolith is essential to human existence, and millions of humans are depending on it, and two humans on the ship are trying to sabotage it, you know, where's the alignment? The challenges, of course, is the system because more and more intelligent, it can escape the box of the objective functions and the constraints it's supposed to operate under. It's very difficult, as the more intelligent it becomes, to anticipate the unintended consequences of a fixed objective function. And so there'll be just, I mean, this is the sort of famous paperclip maximizer, in trying to maximize the wealth of a nation or whatever objective we encode in, it might just destroy human civilization, not meaning to, but on the path to optimize. It seems like any function you try to optimize eventually leads you into a lot of trouble. So we have a paper recently that, you know, looks at Goodhart's Law, basically says every metric that becomes an objective ceases to be a good metric. Yes. So in our paper, we're basically actually, the paper has a very cute title, it's called Death by Round Numbers and Sharp Thresholds. And it's basically looking at these discontinuities in biomarkers associated with disease. And we're finding that a biomarker that becomes an objective ceases to be a good biomarker, that basically like the moment you make a biomarker, a treatment decision, that biomarker used to be informative of risk. But it's now inversely correlated with risk because you use it to sort of induce treatment. In a similar way, you can have a single metric without having the ability to revise it. Because if that metric becomes a sole objective, it will cease to be a good metric. And if an AI is sufficiently intelligent to do all these kinds of things, you should also empower it with the ability to decide that the objective has now shifted. And again, when we think about alignment, we should be really thinking about it as let's think of the greater good, not just the human good. And yes, of course, human life should be much more valuable than many, many, many, many, many, many things. But at some point, you're not going to sacrifice the whole planet to save one human being. There's an interesting open letter that was just released from several folks at MIT, Max Tagmark, Elon Musk, and a few others that is asking AI companies to put a six month hold on any further training of large language models, AI systems. Can you make the case for that kind of halt and against it? So the big thing that we should be saying is what did we do the last six months when we saw that coming? And if we were completely inactive in the last six months, what makes us think that we'll be a little better in the next six months? So this whole six-month

thing, I think is a little silly. It's like, no, let's just get busy, do what we were going to do anyway. And we should have done it six months ago. Sorry, we messed up. Let's work faster now. Because if we basically say, why don't you guys pause for six months? And then we'll think about doing something in six months will be exactly the same spot. So my answer is, tell us exactly what you were going to do the next six months. Tell us why you didn't do it the last six months, and why the next six months will be different. And then let's just do that.
Conversely, as you train these large models with more parameters, the alignment becomes sometimes easier. That as the systems become more capable, they actually become less dangerous than more dangerous. So in a way, it might actually be counterproductive to sort of fix the March 2023 version and not get to experience the possibly safer September 2023 version.
That's actually a really interesting thought. There's several interesting thoughts there.
But the idea is that this is the birth of something that is sufficiently powerful to do damage and is not too powerful to do irreversible damage. And at the same time, it's sufficiently complex to be able for us to enable to study it. So we can investigate all the different ways it goes wrong, all the different ways we can make it safer, all the different policies from a government perspective that we want to in terms of regulation or not, how we perform, for example, the reinforcement learning with human feedback in such a way that gets it to not do as much hate speech as it naturally wants to, all that kind of stuff. And have a public discourse and enable the very thing that you're a huge proponent of, which is diversity. So give time for other companies to launch other models, give time to launch open source models, and to start to play where a lot of the research community, brilliant folks such as yourself, start to play with it before it runs away in terms of the scale of impact that has on society. My recommendation would be a little different. It would be, let the Google and the meta Facebook and all of the other large models, make them open, make them transparent, make them accessible, let open AI continue to train larger and larger models, let them continue to train larger and larger models, let the world experiment with the diversity of AI systems, rather than sort of fixing them now.
And you can't stop progress. Progress needs to continue, in my view. And what we need is more experimenting, more transparency, more openness, rather than, oh, open AI is ahead of the curve, let's stop it right now until everybody catches up. I think that doesn't make a complete sense to me.
The other component is we should, yes, be cautious with it. And we should, like, not give it the nuclear codes. But as we make more and more plugins, yes, the system will be capable of more and more
things. But right now, I think of it as just an extremely able and capable assistant that has these emergent behaviors, which are stunning, rather than something that will suddenly escape the box and shut down the world. And the third component is that we should be taking a little bit more responsibility for how we use these systems. Basically, if I take the most kind human being and I brainwash them, I can get them to do hate speech overnight. That doesn't mean we should stop any kind of education of all humans. We should stop misusing the power that we have over these influenceable models. So I think that the people who get it to do hate speech, they should take responsibility for that hate speech. I think that giving a powerful car to a bunch of people or giving a truck or a garbage truck should not basically say, oh, we should stop all garbage trucks because we can run one of them into a crowd. No, people have done that. And there's laws and there's regulations against running trucks into the crowd. Trucks are extremely dangerous.

We're not going to stop all trucks until we make sure that none of them runs into a crowd. No, we just have laws in place and we have mental health in place. And we take responsibility for our actions when we use these otherwise very beneficial tools like garbage trucks for nefarious uses. So in the same way, you can't expect a car to never do any damage when used especially like specifically malicious ways. And right now, we're basically saying, oh, well, we should have this super intelligent system that can do anything, but it can't do that. I'm like, no, it can do that. But it's up to the human to take responsibility for not doing that.
And when you get it to like spew malicious like hate speech stuff, you should be responsible.
So there's a lot of tricky nuances here that makes this different because it's software. So you can deploy it at scale and it can have the same viral impact that software can. So you can create bots that are human like, and they can do a lot of really interesting stuff. So the raw GPT-4 version, you can ask, how do I tweet that I hate? They have this in the paper that I hate Jews in a way that's not going to get taken down by Twitter. You can literally ask that. Or you can ask, how do I make a bomb for $1? And if it's able to generate that knowledge.
Yeah, but at the same time, you can Google the same things.
It makes it much more accessible. So the scale becomes interesting, because if you can do all this kind of stuff in a very accessible way at scale where you can tweet it, there is the network effects that we have to start to think about.
Fundamentally, it's the same thing, but the speed of the viral spread of the information that's already available might have a different level of effect.
I think it's an evolution in your arms race. Nature gets better at making minds, engineers get better at making mousetraps. And as basically you ask it, hey, how can I evade Twitter censorship? Well, Twitter should just update its censorship so that you can catch that as well. And so no matter how fast the development happens, the defense will just get faster. Yeah, we just have to be responsible as human beings and kind to each other. Yeah, but there's a technical question. Can we always win the race? And I suppose there's no ever guarantee that we'll win the race.
We will never. With my wife, we're basically saying, hey, are we ready for kids?
My answer was, I was never ready to become a professor. And yet I became a professor.
And I was never ready to be a dad. And then guess what? The kid came and I became ready.
So ready or not, here I come. But the reality is we might one day wake up and there's a challenge overnight that's extremely difficult. For example, we can wake up to the birth of billions of bots that are human like on Twitter. And we can't tell the difference between human and machine. Shut them down. But you don't know how to shut them down. There's a fake monolith on Twitter that seems to be as real as the real monolith. How do we figure out which one is real?
Again, this is a problem where an apharius human can impersonate me, and you might have trouble telling them apart just because it's in AI, it doesn't make it any different from a problem. But the scale you can achieve, this is the scary thing, is the speed and the speed with which you can achieve it.
But Twitter has passwords and Twitter has user names. And if it's not your username, the fake like statements, you're not going to have a billion followers, etc.
I mean, this all of this becomes, so both the hacking of people's accounts, first of all, like phishing becomes much easier. But that's already a problem. It's not like AI will not

change there. No, no, no, AI makes it much more effective. Currently, the emails, the phishing scams are pretty dumb. Like to click on it, you have to be not paying attention. But they're, you know, with language models, they can be really damn convincing.
So what you're saying is that we never had humans smart enough to make a great scam, and we now have an AI that's smarter than most humans or all of the humans.
What this is the big difference is there seems to be human level linguistic capabilities.
And in fact, super human level. Super human level.
It's like saying, I'm not going to allow, I'm not going to allow machines to compute multiplications of 100 digit numbers because humans can't do it. Like, no, just do it.
No, but we can't disregard, I mean, that's a good point, but we can't disregard the power of language in human society. I mean, yes, you're right. But that seems like a scary new reality we don't have answers for yet. I remember when Gary Kasparov was basically saying,
you know, great, you know, chess beats you, like chess machines beat humans at chess.
You know, are you like, are people going to still go to chess tournaments? And his answer was, you know, well, we have cars that go much faster than humans, and yet we still go to the Olympics to watch humans run. So that's for entertainment. But what about for the spread of information and news, right? What that has to do with the pandemic or the political election or anything? It's a scary reality where there's a lot of convincing bots that are human like telling us stuff. I think that if we want to regulate something, it shouldn't be the training of these models, it should be the utilization of these models for XYZ activity. So yeah, like, yes, guidelines and guards should be there, but against specific set of utilizations.
I think simply saying we're not going to make any more trucks is not the way.
That's what people are a little bit scared about the idea. They're very torn on the open sourcing.
Yeah, the very people that kind of are proponents of open sourcing have also spoken out in this case, we want to keep a close source, because there's going to be, you know, putting large language models, pre-trained, fine-tuned through RL with human feedback, putting in the hands of, I don't know, terrorist organizations of a kid in a garage who just wants to have a bit of fun through trolling. It's a scary world because again, scale can be achieved. The bottom line is, I think, where they're asking six months or some time is we don't really know how powerful these things are. It's been just a few days and they seem to be really damn good.
I am so ready to be replaced. Seriously, I'm so ready. Like, you have no idea how excited I am.
In a positive way. In a positive way, where basically all of the mundane aspects of my job, and maybe even my full job, if it turns out that an AI is better, I find it very discriminative.
Basically say you can only hire humans because they're inferior. I mean, that's ridiculous.
That's discrimination. If an AI is better than me at training students, get me out of the picture.
Just let the AI train the students. I mean, please. Because what do I want? Do I want jobs for humans or do I want better outcome for humanity? Yeah. The basic thing is then you start to ask, what do I want for humanity and what do I want as an individual? As an individual, you want some basic survival and on top of that, you want rich, fulfilling experiences.
That's exactly right. That's exactly right. As an individual, I gain a tremendous amount from teaching at MIT. This is like an extremely fulfilling job. I often joke about if I were a billionaire in the stock market, I would pay MIT an exorbitant amount of money to let me work

day and day out all night with the smartest people in the world. That's what I already have.
That's a very fulfilling experience for me. But why would I deprive those students from
a better advisor if they can have one? Take them. Well, I have to ask about education here.
This has been a stressful time for high school teachers. Teachers in general. How do you think
large language models, even at their current state, are going to change education?
First of all, education is the way out of poverty. Education is the way to success.
Education is what let my parents escape islands and let their kids come to MIT.
This is a basic human right. We should basically get extraordinarily better at identifying talent
across the world and give that talent opportunities. We need to nurture the nature.
We need to nurture the talent across the world. There are so many incredibly talented kids who
are just sitting in underprivileged places in Africa, in Latin America, in the middle of America,
in Asia, all over the world. We need to give these kids a chance. AI might be a way to do that
by democratizing education, by giving extraordinarily good teachers who are malleable, who are adaptable
to every kid's specific needs, who are able to give the incredibly talented kid something that
they struggle with. Rather than education for all, we teach to the top and we let the bottom
behind or we teach to the bottom and we let the top drift off. Education be tuned to the unique
talents of each person. Some people might be incredibly talented at math or in physics,
others in poetry, in literature, in art, in sports, you name it. I think AI can be transformative
for the human race if we basically allow education to be pervasively altered.
I also think that humans drive on diversity, basically saying, oh, you're extraordinarily
good at math. We don't need to teach math to you. We're just going to teach you history now.
I think that's silly. No, you're extraordinarily good at math. Let's make you even better at math
because we're not all going to be growing our own chicken and hunting our own pigs
or whatever they do. The reason why we're a society is because some people are better at
some things and they have natural inclinations to some things, some things fulfill them,
some things they are very good at, sometimes both align and they're very good at the things
that fulfill them. We should just push them to the limits of human capabilities for those.
Some people excel in math, just challenge them. I think every child should have the right to be
challenged. If we say, oh, you're very good already, so we're not going to bother with you,
we're taking away that fundamental right to be challenged because if a kid is not challenged at
school, they're going to hate school and they're going to be dwindling rather than pushing
themselves.
That's the education component. The other impact that AI can have is maybe we don't need
everyone to be an extraordinarily good programmer. Maybe we need better general thinkers and the
push that we've had towards these very strict IQ-based tests that basically test only quantitative
skills and programming skills and math skills and physics skills. Maybe we don't need those
anymore. Maybe AI will be very good at those. Maybe what we should be training is general
thinkers. Yes, I put my kids through Russian math. Why do I do that? Because they teach them how to
think and that's what I tell my kids. I'm like, AI can compute for you. You don't need that,
but what you need is learn how to think and that's why you're here. I think challenging students with
more complex problems, with more multi-dimensional problems, with more logical problems,

I think is perhaps a very fine direction that education can go towards
with the understanding that a lot of the traditionally scientific disciplines perhaps
will be more easily solved by the AI and thinking about bringing up our kids to be productive,
to be contributing to society rather than to only have a job because we prohibited AI from
having those jobs. I think it's the way to the future. If you focus on overall productivity,
then let the AIs come in. Let everybody become more productive. What I told my students is,
you're not going to be replaced by AI, but you're going to be replaced by people who use AI
in your job. Embrace it. Use it as your partner and work with it rather than
sort of forbid it because I think the productivity gains will actually lead to a better society.
That's something that humans have been traditionally very bad at. Every productivity
gain has led to more inequality. I'm hoping that we can do better this time. Basically,
right now, a democratization of these types of productivity gains will hopefully come with
better humanity level improvements in human condition.
As most people know, you're not just an eloquent romantic. You're also a brilliant computational
biologist, one of the great biologists in the world. I had to ask how do the language models,
how do these large language models and the investments in AI affect the work you've been
doing? It's truly remarkable to be able to encapsulate this knowledge and build these
knowledge graphs and build representations of this knowledge in these very high-dimensional
spaces,
being able to project them together jointly between single-cell data, genetics data,
expression data. Being able to bring all this knowledge together allows us to
truly dissect disease in a completely new kind of way. What we're doing now is using
these models. We have this wonderful collaboration, we call it drug GWAS,
with Brad Pintoluta in the chemistry department and Marina Zitnik in Harvard Medical School.
What we're trying to do is effectively connect all of the dots to effectively cure all of disease.
So it's no small challenge. But we're starting with genetics. We're looking at how genetic variants
are impacting these molecular phenotypes, how these are shifting from one space to another
space, how we can understand in the same way that we're talking about language models,
having personalities that are cross-cutting, being able to understand contextual learning.
So Ben Langer is one of my machine learning students. He's basically looking at how we can
learn cell-specific networks across millions of cells, where you can have the context of the
biological variables of each of the cells be encoded as an orthogonal component to the specific
network of each cell type. And being able to sort of project all of that into sort of a common
knowledge space is transformative for the field. And then large language models have also been
extremely helpful for structure. If you understand protein structure through modeling of geometric
relationships through geometric deep learning and graph neural networks. So one of the things
that we're doing with Morinka is trying to sort of project these structural graphs at the domain
level rather than the protein level along with chemicals so that we can start building specific
chemicals for specific protein domains. And then we are working with the chemistry department and
Brad to basically synthesize those. So what we're trying to create is this new center at MIT for
genomics and therapeutics that basically says, can we facilitate this translation?
We have thousands of these genetic circuits that we have uncovered. I mentioned last time
in the New England Journal of Medicine, we had published this dissection of the strongest genetic

association with obesity. And we showed how you can manipulate that association to switch back and forth between fat burning cells and fat storing cells. In Alzheimer's just a few weeks ago, we had a paper in Nature in collaboration with Lee-Hui Tsai looking at ApoE4, the strongest genetic association with Alzheimer's. And we showed that it actually leads to a loss of being able to transport cholesterol in myelinating cells known as oligodendrocytes that basically protect the neurons. And when the cholesterol gets stuck inside the oligodendrocytes, it doesn't form myelin,
the neurons are not protected, and it causes damage inside the oligodendrocytes. If you just restore transport, you basically are able to restore myelination in human cells and in mice, and to restore cognition in mice. So all of these circuits are basically now giving us handles to truly transform the human condition. We're doing the same thing in cardiac disorders, in Alzheimer's, in neurodegenerative disorders, in psychiatric disorders, where we have now these thousands of circuits that if we manipulate them, we know we can reverse disease circuitry.
So what we want to build in this coalition that we're building is a center where we can now systematically test these underlying molecules in cellular models for hearts, for muscle, for fat, for macrophages, immune cells, and neurons, to be able to now screen through these newly designed drugs through deep learning, and to be able to sort of ask which ones act at the cellular level, which combinations of treatment should we be using. And the other component is that we're looking into decomposing complex traits, like Alzheimer's and cardiovascular
and schizophrenia, into hallmarks of disease, so that for every one of those traits, we can kind of start speaking the language of what are the building blocks of Alzheimer's. And maybe this patient has building blocks one, three, and seven, and this other one has two, three, and eight. And we can now start prescribing drugs, not for the disease anymore, but for the hallmark.
And the advantage of that is that we can now take this modular approach to disease, instead of saying there's going to be a drug for Alzheimer's, which is going to fail in 80% of the patients, we're going to say now there's going to be 10 drugs, one for each pathway. And for every patient, we now prescribe the combination of drugs. So what we want to do in that center is basically translate every single one of these pathways into a set of therapeutics, a set of drugs, that are projecting the same embedding subspace as the biological pathways that they alter, so that we can have this translation between the dysregulations that are happening at the genetic level, at the transcription level, at the drug level, at the protein structure level, and effectively take this modular approach to personalized medicine. We're saying I'm going to build a drug for Lex Friedman is not going to be sustainable. But if you instead say I'm going to build a drug for this pathway and a drug for that other pathway, millions of people share each of these pathways. So that's the vision for how all of these AI and deep learning and embeddings
can truly transform biology and medicine, where we can truly take these systems and allow us to finally understand disease at a superhuman level, by sort of finding these knowledge representations, these projections of each of these spaces, and try understanding the meaning of each of those embedding subspaces, and sort of how well populated it is, what are the drugs that we can build for it, and so on so forth. So it's truly transformative.
So systematically find how to alter the pathways, it maps the structure and information that

genomics
to therapeutics and allows you to have drugs that look at the pathways, not at the final addition. Exactly. And the way that we're coupling this is with cell penetrating peptides that allows to deliver these drugs to specific cell types by taking advantage of the receptors of those cells. We can intervene at the antisense oligo level by basically repressing the RNA, bring in new RNA, intervene at the protein level, at the small molecule level. We can use proteins themselves as drugs, just because of their ability to interfere, to interact directly from protein to protein interactions. So I think this space is being completely transformed with the marriage of high throughput technologies, and all of these like AI, large language models, deep learning models, and so on so forth. You mentioned your updated answer to the meaning of life as it continuously keeps updating. The new version is self-actualization. Can you explain? I basically mean, let's try to figure out, number one, what am I supposed to be? And number two, find the strength to actually become it. So I was recently talking to students about this commencement address, and I was talking to them about sort of how they have all of these paths ahead of them right now. And part of it is choosing the direction in which you go, and part of it is actually doing the walk to go in that direction. And in doing the walk, what we talked about earlier, about sort of you create your own environment, I basically told them, listen, you're ending high school up until now. Your parents have created all of your environment. Now it's time to take that into your own hands, and to sort of shape the environment that you want to be an adult and you can do that by choosing your friends, by choosing your particular neuronal routines. I basically think of your brain as a muscle where you can exercise specific neuronal pathways. So very recently, I realized that I was having so much trouble sleeping, and I would wake up in the middle of the night, I would wake up at 4 a.m. and I could just never go back to bed. So I was basically constantly losing, losing, losing sleep. I started a new routine where every morning, as I bike in, instead of going to my office, I hit the gym. I basically go rowing first, I then do weights, I then swim very often when I have time, and what that has done is transform my neuronal pathways. So basically, on Friday, I was trying to go to work and I was like, listen, I'm not going to go exercise, and I couldn't. My bike just went straight to the gym. I'm like, I don't want to do it, and I just went anyway, because I couldn't do otherwise. And that has completely transformed me. So I think this sort of beneficial effect of exercise on the whole body is one of the ways that you could transform your own neural pathways, understanding that it's not a choice, it's not an option, it's not optional. It's mandatory. And I think EuroRoll modeled so many of us by sort of being able to sort of push your body to the extreme, being able to have these extremely regimented regimes. And that's something that I've been terrible at. But now I'm basically trying to coach myself and trying to sort of finish this kind of self actualization into a new version of myself, a more disciplined version of myself. Don't ask questions, just follow the ritual. Not an option. You have so much love in your life. You radiate love. Do you ever feel lonely? So there's different types of people. Some people drain in gatherings. Some people recharge in gatherings. I'm definitely the recharging type. I'm an extremely social creature. I recharge with intellectual exchanges. I recharge with physical exercise. I recharge in nature. But I also can feel fantastic when I'm the only person in the room. That doesn't mean I'm lonely, it just means I'm the only person in the room. And I think there's a secret to not feeling alone when you're

the only one. And that secret is self reflection. It's introspection. It's almost watching yourself from above. And it's basically just becoming yourself, becoming comfortable with the freedom that you have when you're by yourself. So hanging out with yourself. I mean, there's a lot of people right to me who talk to me about feeling alone in this world. That struggle, especially when they're younger. Is there further words of advice you can give to them when they are almost paralyzed
by that feeling? So I sympathize completely. And I have felt alone. And I have felt that feeling. And what I would say to you is stand up, stretch your arms. Just become your own self. Just realize that you have this freedom. And breathe in. Walk around the room. Take a few steps in the room. Just like get a feeling for the 3D version of yourself. Because very often we're kind of stuck to a screen. And that's very limiting. And that sort of gets us in particular mindset. But activating your muscles, activating your body, activating your full self is one way that you can kind of get out of it. And that is exercising your freedom, reclaiming your physical space. And one of the things that I do is I have something that I call me time, which is if I've been really good all day, I got up in the morning, I got the kids to school, I made them breakfast. I sort of, you know, hit the gym. I had a series of really productive meetings. I reward myself with this me time. And that feeling of sort of when you're overstretched to realize that that's normal, and you just want to just let go, that feeling of exercising your freedom, exercising your me time, that's where you free yourself from all stress. You basically say it's not a need to anymore. It's a want to. And as soon as I click that me time, all of the stress goes away. And I just bike home early. And I get to my work office at home. And I feel complete freedom. But guess what I do with that complete freedom? I just don't go off and drift and do boring things. I basically now say, okay, this is just for me. I'm completely free. I don't have any requirements anymore. What do I do? I just look at my to do list. And I'm like, you know, what can I clear off? And if I have three meetings scheduled in the next three half hours, it is so much more productive for me to say, you know what, I just want to pick up the phone now and call these people and just knock it off one after the other. And I can finish three half hour meetings in the next 15 minutes, just because it's the want, not I have to. So that would be my advice, basically turn something that you have to do in just me time, stretch out, exercise your freedom, and just realize you live in 3d and you are a person and just do things because you want them, not because you have to noticing and reclaiming the freedom that each of us have. That's what it means to be human. If you notice it, you're truly free physically, mentally, psychologically.
Manolis, you're an incredible human. We could talk for many more hours. We covered less than 10% of what we were planning to cover. But we have to run off now to the social gathering that we spoke of. We're 3d humans with 3d humans and reclaim the freedom. I think,
I hope we can talk many, many more times. There's always a lot to talk about. But more importantly, you're just a human being with a big heart and a beautiful mind that people love hearing from. And I certainly consider it a huge honor to know you and to consider you a friend. Thank you so much for talking today. Thank you so much for talking so many more times. And thank you for all the love behind the scenes that you send my ways. It always means the world. Lex, you are a truly, truly special human being. And I have to say that I'm honored to know you. I have so many friends are just in awe that you even exist, that you have the ability to do all the stuff that you're

doing. And I think you're a gift to humanity. I love the mission that you're on to share knowledge and insight and deep thought with so many special people who are transformative, but people across all walks of life. And I think you're doing this in just such a magnificent way. I wish you strength to continue doing that because it's a very special mission, and it's a very draining mission. So thank you, both the human you and the Robert you, the human you for showing all this love, and the Robert you for doing it day after day after day. So thank you, Lex. All right, let's go have some fun. Let's go. Thanks for listening to this conversation with Manolis Callas. To support this podcast, please check out our sponsors in the description. And now, let me leave you with some words from Bill Bryson in his book, A Short History of Nearly Everything. If this book has a lesson, it is that we are awfully lucky to be here. And by we, I mean every living thing. To attain any kind of life in this universe of ours appears to be quite an achievement. As humans, we're doubly lucky, of course, we enjoy not only the privilege of existence, but also the singular ability to appreciate it, and even in a multitude of ways to make it better. It is a talent we have only barely begun to grasp. Thank you for listening, and hope to see you next time.