The following is a conversation with Yoshabak his third time on this podcast.
Yosha is one of the most brilliant and fascinating minds in the world
exploring the nature of intelligence consciousness and computation.
And he's one of my favorite humans to talk to about pretty much anything and everything.
And now a quick few second mention of his sponsor.
Check them out in the description.
It's the best way to support this podcast.
We've got Numeri for the world's hardest data science tournament.
Aidsleep for naps.
Masterclass for learning and AG1 for health.
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, but if you must skip them,
please still check out our sponsors.
I enjoy their stuff.
Maybe you will too.
This show is brought to you by Numeri.
A hedge fund that uses AI and machine learning to make investment decisions
is basically a super difficult machine learning tournament
that uses real data and people's submitted models that try to predict the market.
I love difficult real world data sets.
You may know that for a long time and still I've been interested in real world robotics.
One of the largest scale deployment of real world robotics is autonomous vehicles.
Autonomous driving and semi-autonomous driving, the stakes are very high.
The same is true for financial markets.
And so it's really interesting that Numeri presents to you the real world data
of financial markets and presents you an easy accessible mechanism
by which to test, deploy, and compete with others in this kind of data set.
So it's a really great way if you're interested in data science
and machine learning to learn, to compete, to have fun, all that kind of stuff.
Head over to numeri.ai slash lex to sign up for a tournament
and whole new machine learning skills.
That's numeri.ai slash lex for a chance to play against me
and win a share of the tournament's prize pool.
This episode is also brought to you by 8Sleep and its new Pi 3 mattress.
In scorching Texas heat, the thing I go to to escape, to escape nature,
or the external harsh conditions of nature, and going to the nature of my own mind.
Wherever that weird and beautiful dream world is, the place that has no rules, no boundaries,
no limits, no physics, no constraints on what is possible and what is impossible.
The dream world that we go to, what is that world?

It's the same world as imagination.
It's such a fascinating world.
The human mind, its capabilities are just so incredibly fascinating
and one of the ways to explore that is the dream.
But it's the return from the dream world that is the most refreshing to me.
That's why I love naps.
It's a quick stroll through the dream world in your back
and taking on the challenges of the day in here and now.
Anyway, if you're into naps as much as me, you should check out 8Sleep
and you'll get special savings when you go to 8Sleep.com slash Lex.
This show is also brought to you by Masterclass.
$10 a month gets you an all-access pass to watch courses from the best people in the world
and their respective disciplines.
The list of courses I've personally watched and enjoyed just lasts forever, but they have
everybody and anybody you ever want to listen to.
I'll listen to Martin Scorsese, Tony Hogg, Jane Goodall, Neil Gaiman, Daniel Negron,
before I interviewed him, Gary Kasparov, Carlos Santana, Will Wright,
Neil deGrasse, Tyson, Chris Hadfield.
The list is incredible.
I'm a huge believer that learning about a thing,
at least part of learning about a thing, should involve learning or listening
to the best people in the world at that thing.
It's not only the advice they give.
It's not only the analysis or the description of how they approach the thing,
but in the way they see life, in the way they carry themselves physically and mentally.
You get to watch mastery and it's so beautiful that human beings are able to reach
at the very top of excellence and sometimes break through the boundaries,
the limits of what was thought possible before, and it's just beautiful to watch those humans.
It's beautiful, it's inspiring.
It's great to celebrate that.
It's great to learn from that.
Anyway, get unlimited access to every masterclass and get 15% off
an annual membership at masterclass.com slash lex.
That's masterclass.com slash lex.
This show is brought to you by AG1.
They're all in one daily drink, brings happiness to me.
And daily, for me, is twice daily.
It brings happiness, health, and ensures that all the crazy physical and mental stuff I do
is built on a foundation of basic nutritional health.
It's the super multivitamin that I use.
It also is one of the components of daily habits that I have in my life.
And so whenever I do this thing, I feel grounded.
I feel happy.

I feel like I have my life together.
So you could do that both at home and with the travel packs when you're traveling.
In fact, it's one of the things that makes me feel like I'm at home when I'm traveling.
I'll drink an AG1 and it'll feel good.
It'll put a smile on my face.
It's green, it tastes delicious.
What else do you want?
They'll give you a one month supply of fish oil when you sign up at drinkag1.com slash lex.
This is the Lex Friedman podcast.
The supported.
Please check out our sponsors in the description.
And now, dear friends, here's Yosha Bach.
You wrote a post about levels of lucidity.
Quote, as we grow older, it becomes apparent that our self-reflexive mind is not just gradually accumulating ideas about itself, but that it progresses in somewhat distinct stages.
So there's seven of the stages.
Stage one, reactive survival infant.
Stage two, personal self, young child.
Stage three, social self, adolescence, domesticated adult.
Stage four is rational agency, self-direction.
Stage five is self-authoring.
That's full adult.
You've achieved wisdom, but there's two more stages.
Stage six is enlightenment.
Stage seven is transcendence.
Can you explain each or the interesting parts of each of these stages?
And what's your sense, why there are stages of this, of lucidity as we progress through life in this too short life?
This model is derived from a concept by the psychologist Robert Keegan.
And he talks about the development of the self as a process that happens in principle by some kind of reverse engineering of a mind where you gradually become aware of yourself and thereby build structure that allows you to interact deeper with the world and yourself.
And I found myself using this model not so much as a developmental model.
I'm not even sure if it's a very good developmental model because I saw my children not progressing exactly like that.
And I also suspect that you don't go through these stages necessarily in succession.
And it's not that you work through one stage and then you get into the next one.
Sometimes you revisit them.
Sometimes stuff is happening in parallel.
But it's, I think, a useful framework to look at what's present in the structure of a person and how they interact with the world and how they relate to themselves.
So it's more like a philosophical framework that allows you to talk about how minds work.
And at first, when we are born, we don't have a personal self yet, I think.

Instead, we have an attentional self.
And this attentional self is initially, in the infant, tasked with building a world model and also an initial model of the self.
But mostly, it's building a game engine in the brain that is tracking sensory data and uses it to explain it.
And in some sense, you could compare it to a game engine like Minecraft or so, so colors and sounds.
People are all not physical objects.
They are creation of our mind at a certain level of coarse graining.
Models that are mathematical, that use geometry and that use manipulation of objects and so on to create scenes in which we can find ourselves and interact with them.
So Minecraft.
Yeah.
And this personal self is something that is more or less created after the world is finished, after it's trained into the system, after it has been constructed.
And this personal self is an agent that interacts with the outside world.
And the outside world is not the world of quantum mechanics, not the physical universe, but it's the model that has been generated in our own mind.
Right?
And this is us.
And we experience ourselves interacting with that outside world that is created on the side of our own mind.
And outside of our self, there are feelings and they presented our interface to this outside world.
They pose problems to us.
These feelings are basically attitudes that our mind is computing, that tell us what's needed in the world, the things that we are drawn to, the things that we are afraid of.
And we are tasked with solving this problem of satisfying the needs, avoiding the aversions, following on our inner commitments and so on, and also modeling ourselves and building the next stage.
So after we have this personal self and Stage 2 online, many people form a social self.
And this social self allows the individual to experience themselves as part of a group.
It's basically this thing that when you are playing in a team, for instance, you don't notice yourself just as a single note that is reaching out into the world, but you're also looking down.
You're looking down from this entire group and you see how this group is looking at this individual and everybody in the group is in some sense emulating this group spirit to some degree.
And in this state, people are forming their opinions by assimilating them from this group mind, basically gain the ability to act a little bit like a hive mind.
But are you also modeling the interaction of how opinion shapes and forms through the interaction of the individual nodes within the group?
Yeah, it's basically the way in which people do it in a stage is that they experience what are the opinions of my environment.

They experience the relationship that they have to their environment
and they resonate with people around them and get more opinions through this interaction,
the way in which they relate to others.
And at Stage 4, you basically understand that stuff is true and false independently
about other people belief and you have agency over your own beliefs in that stage.
You basically discover epistemology, the rules about determining what's true and false.
You start to learn how to think.
Yes. I mean, at some level, you're always thinking you are constructing things and I
believe that this ability to reason about your mental representation is what we mean by thinking.
It's an intrinsically reflexive process that requires consciousness.
Without consciousness, you cannot think.
You can generate the content of feelings and so on outside of consciousness.
It's very hard to be conscious of how your feelings emerge,
at least in the early stages of development, but thoughts is something that you always control.
And if you are inert like me, you often have to skip Stage 3 because you lack the intuitive
empathy with others because in order to resonate with a group, you need to have a quite similar
architecture. And if people are via differently, then it's hard for them to resonate with other
people and basically have empathy, which is not the same as compassion, but it is a shared
perceptual mental state. Empathy happens not just via inference about the mental states of others,
but it's a perception of what other people feel and where they're at.
Can't you not have empathy while also not having a similar architecture, cognitive
architecture as the others in the group? I think yes, but I experienced that too,
but you need to build something that is like a meta-architecture. You need to
be able to embrace the architecture of the other to some degree or find some common ground.
It's also this issue that if you are inert, normally, as in your typical people,
have difficulty to resonate with you. And as a result, they have difficulty understanding you
unless they have enough wisdom to feel what's going on there.
Isn't the whole process of the Stage 3 is to figure out the API to the other humans that
have different architecture and you yourself publish public documentation for the API
that people can interact with for you? Isn't this the whole process of socializing?
My experience as a child growing up was that I did not find any way to interface with the Stage 3
people and they didn't do that with me. Did you try?
Yeah, of course, I tried it very hard, but it was only when I entered a mathematics school
at 9th grade, lots of other nerds were present that I found people that I could deeply resonate with
and had the impression that yes, I have friends now, I found my own people. And before that,
I felt extremely lonely in the world. There was basically nobody I could connect to.
And I remember there was one moment in all these years where there was a school exchange
and it was a Russian boy, a kid from the Russian Garnison in the station in Eastern
Germany who visited our school and we played a game of chess against each other.
And we looked into each other's eyes and we sat there for two hours playing this game of chess
and I had the impression this is a human being. He understands what I understand.
And we didn't even speak the same language.
I wonder if your life could have been different if you knew that it's okay to be different,

to have a different architecture. Whether accepting the interface is hard to figure out, takes a long time to figure out and it's okay to be different. In fact, it's beautiful to be different. It was not my main concern. My main concern was mostly that it was alone.
It was not so much the question, is it okay to be the way? I couldn't do much about it, so I had to deal with it. But my main issue was that I was not sure if I would ever meet anybody growing up that I would connect to at such a deep level that I would feel that I could belong.
So there's a visceral, undeniable feeling of being alone.
And I noticed the same thing when I came into the mass school that I think at least half, probably two-thirds of these kids were severely traumatized as children growing up and in large part due to being alone because they couldn't find anybody to relate to.
Don't you think everybody's alone, deep down? No.
All right. I'm not alone anymore. It took me some time to update and to get over the trauma and so on, but I felt that in my 20s, I had lots of friends and I had my place in the world and I had no longer doubts that I would never be alone again.
Is there some aspect to which we're alone together? You don't see a deep loneliness inside yourself still? No. Sorry.
Okay. So that's the nonlinear progression through the stages, I suppose. You caught up on stage three at some point. So we are at stage four and so basically I find that many nerds jump straight into stage four by passing stage three. Do they return to it then later?
Yeah. Of course, sometimes they do, not always. The question is basically, do you stay a little bit autistic or do you catch up? And I believe you can catch up. You can build this missing structure and basically experience yourself as part of a group, learn intuitive empathy and develop the sense, this perceptual sense of feeling what other people feel. And before that, I could only basically feel this when I was deeply in love with somebody and we synced.
So there's a lot of friction to feeling that way. It only was certain people as opposed to it comes naturally as frictionless. But this is something that basically later I felt started to resolve itself for me to a large degree. What was the trick?
In many ways, growing up and paying attention. Meditation did tap. I had some very crucial experiences in getting close to people, building connections, cuddling a lot in my student years.
So really paying attention to the, what is it, to the feeling another human being fully?
Loving other people and being loved by other people and building a space in which you can be safe and can experiment and touch a lot and be close to somebody a lot. And that over time, basically, at some point you realize, oh, it's no longer that I feel locked out, but I feel connected and I experience where somebody else is at. And normally my mind is racing very fast at high frequencies. So it's not always working like this. Sometimes it works better, sometimes it works less. But also don't see this as a pressure. It's more, it's interesting to observe myself, which frequency I'm at and at which mode somebody else is at.
Yeah, man, the mind is so beautiful in that way. Sometimes it comes so natural to me, so easy to pay attention, pay attention to the world fully to other people fully.
And sometimes the stress over silly things is overwhelming. It's so interesting that the mind is that roller coaster in that way. At stage five, you discover how identity is constructed. Self-offering. Realize that your values are not terminal, but they are instrumental to achieving world that you like and aesthetics that you prefer. And the more you understand this,

the more you get agency over how your identity is constructed. And you realize that identity and interpersonal interaction is a costume. And you should be able to have agency over that costume, right? It's useful to be a costume. It tells something to others and it allows to interface and roles. But being locked into this is a big limitation.
The word costume kind of implies that it's fraudulent in some way. Is costume a good word for you? Like to present ourselves to the world? In some sense, I learned a lot about costumes at Burning Man. Before that, I did not really appreciate costumes and saw them more as uniforms, like wearing a suit if you are working in a bank or if you are trying to get startup funding from a VC in Switzerland, right? Then you dress up in a particular way. And this is mostly to show the other side that you are willing to play by the rules and you understand what the rules are. But there is something deeper. When you are at Burning Man, your costume becomes self-expression and there is no boundary to the self-expression. You're basically free to wear what you want to express other people what you feel like this day and what kind of interactions you want to have. Is the costume a kind of projection of who you are?
That's very hard to say because the costume also depends on what other people see in the costume. And this depends on the context that the other people understand. So you have to create something if you want to that is legible to the other side and that means something to yourself.
Do we become prisoners of the costume? Because everybody expects us to.
Some people do, but I think that once you realize that you wear a costume at Burning Man, a variety of costumes, realize that you cannot not wear a costume. Basically everything that you wear and present to others is something that is to some degree in addition to what you are deep inside. So this stage, in parentheses, you put full adult comma wisdom. Why is this full adult? Why would you say this is full and why is it wisdom? It does allow you to understand why other people have different identities from yours. And it allows you to understand that the difference between people who vote for different parties and might have very different opinions and different value systems is often the accident of where they are born and what happened after that to them and what traits they got before they were born. And at some point you realize a perspective where you understand that everybody could be you in a different timeline if you just flip those bits. How many costumes do you have? I don't count. More than one? Yeah, of course. How easy is to do costume changes throughout the day? It's just a matter of energy and interest. When you are wearing your pajamas and you switch out of your pajamas into say a work short and pants, you're making a costume change. And if you're putting on a gown, you're making a costume change. And you could do the same with personality? You could if that's what you're into. There are people which have multiple personalities for interaction in multiple worlds. So if somebody works in a store and you put up a storekeeper personality, when you're working, when you're presenting yourself at work, you develop a subpersonality for this. And the social persona for many people is in some sense a puppet that they're playing like a marionette. And if they play this all the time, they might forget that there is something behind this. There's something what it feels like to be in your skin. And I guess it's very helpful if you're able to get back into this. And for me, the other way around is relatively hard. For me, it's pretty hard to learn how to play consistent social roles from it's much easier just to be real.
Or not real, but to have one costume.
No, it's not quite the same. So basically, when you are wearing a costume at Burning Man and say

you are an extraterrestrial prince, there's something where you are expressing in some sense something that's closer to yourself than the way in which you hide yourself behind standard closing when you go out in the city in the default world. And so this costume that you're wearing at Burning Man allows you to express more of yourself. And you have a shorter distance of advertising to people, what kind of person you are, what kind of interaction you would want to have with them. And so you get much earlier into media stress. And I believe it's regrettable that we do not use the opportunities that we have this custom made closing now to weird costumes that are
much more stylish, that are much more custom made that are not necessarily part of a fashion in which
you express which milieu you're part of and how up to date you are. But you also express how you are as an individual and what you want to do today and how you feel today and what you intend to do
about it. Well, isn't it easier now with individual world to explore different costumes? I mean, that's the kind of idea with virtual reality. That's the idea even with Twitter in two-dimensional screens. You can swap all costumes. You could be as weird as you want. It's easier. For Burning Man,
you have to order things. You have to make things. It's more effort to put on.
It's better if you make them yourselves. Sure. But it's just easier to do digitally, right?
It's not about easy. It's about how to get it right. And for me, the first Burning Man experience I got adopted by a bunch of people in Boston who dragged me to Burning Man and we spent a few weekends doing costumes together. And that was an important part of the experience where the camp
bonded that people got to know each other and we basically grew into the experience that we would have later. So the extraterrestrial prince is based on a true story. I can only imagine what that looks like. Yeah, sure. Okay. So, stage six. At some point, you can collapse the division between a personal self and world generator again. And a lot of people get there via meditation or some of them get there via psychedelics, some of them by accident. And you suddenly notice that you are not actually a person, but you are a vessel that can create a person. And the person is still there. You observe that personal self, but you observe the personal self from the outside. And you notice it's a representation. And you might also notice that the world that is being created is a representation. If not, then you might experience that I am the universe. I am the thing that is creating everything. And of course, what you're creating is not quantum mechanics, and the physical universe, what you're creating is this game engine that is updating the world and you're creating your valence, your feelings, and all the people inside of that world,
including the person that you identify with yourself in this world.
Are you creating the game engine or are you noticing the game engine?
You noticed how you're generating the game engine. And I mean, when you are dreaming at night, if you have a lucid dream, you can learn how to do this deliberately. And in principle,
you can also do it during the day. And the reason why we don't get to do this from the beginning and why we don't have agency of our feelings right away is because we would game it before they have the necessary amount of wisdom to deal with creating this dream that we are in. You don't want to get access to cheat codes too quickly. Otherwise, you won't enjoy the game.

So stage five is already pretty rare. And stage six is even more rare. You both basically find this mostly with advanced Buddhist meditators and so on that dropping into the stage and can induce
it at will and spend time in it. So stage five requires a good therapist. Stage six requires a good Buddhist spiritual leader. Yes. For instance, it could be that it's the right thing to do. But it's not that these stages give you scores or levels that you need to advance to. It's not that the next stage is better. You live your life and the more that works best at any given moment. And when your mind decides that you should have a different configuration, then it's building that configuration. And for many people, they stay happily at stage three and experiences themselves as part of groups. And there's nothing wrong with this. And for some people, this doesn't work. And they're forced to build more agency over their rational beliefs than this and construct their norms rationally. And so they go to this level. And stage seven is something that is more or less hypothetical. That would be the stage in which it's basically a transhumanist stage in which you understand how you work and which the mind fully realizes how it's implemented and can also, in principle, enter different modes in which it could be implemented. And that's the stage that as far as I understand is not open to people yet.
Oh, but it is possible to the process of technology.
Yes. And who knows if there are biological agents that are working at different time scales than us that basically become aware of the way in which they're implemented on ecosystems and can change that implementation and have agency over how they're implemented in the world. And what I find interesting about the discussion about AI alignment, that it seems to be following these stages very much. Most people seem to be in stage three also according to Robert Keegan. I think he says that about 85% of people are in stage three and stay there. And if you're in stage three and your opinions are the result of social assimilation, then what you're mostly worried about and the AI is that the AI might have the wrong opinions. So if the AI says something racist or sexist, we are all lost because we will assimilate the wrong opinions from the AI. And so we need to make sure that the AI has the right opinions and the right values and the right structure. And if you're at stage four, that's not your main concern. And so most nerds don't really worry about the algorithmic bias and the model that it picks up because if there's something wrong with this bias, the AI ultimately will prove it. At some point, we'll get it there that it makes mathematical proofs about reality and then it will figure out what's true and what's false. But you're still worried that the AI might turn you into paperclips because it might have the wrong values. So if it's set up with the wrong function that controls its direction in the world, then it might do something that is completely horrible and there's no easy way to fix it. So that's more like a stage four rationalist kind of worry. And if you are at stage five, you're mostly worried that the AI is not going to be enlightened fast enough because you realize that the game is not so much about intelligence but about agency, about the ability to control the future. And the identity is instrumental to this. And if you are human being, I think at some level, you ought to choose your own identity. You should not have somebody else pick the costume for you and then wear it. But instead, you should be mindful about what you want to be in this world. And I think if you are an agent that is fully malevol that can rewrite its own source code like an AI might do at some point, then the identity that you will have is whatever you can be. And in this way, the AI will maybe become everything like a planetary

control system. And if it does that, then if we want to coexist with it, it means that it will have to share purposes with us. So it cannot be a transactional relationship. We will not be able to use reinforcement learning with human feedback to hardwire its values into it. But this has to happen is probably that it's conscious so it can relate to our own mode of existence where an observer is observing itself in real time and there's a certain temple of frames. And the other thing is that it probably needs to have some kind of transcendental orientation building shared agency.

And in the same way as we do when we are able to enter up with each other into non-transactional relationships. And I find that something that because the stage five is so rare is missing in much of the discourse. And I think that we need in some sense, focus on how to formalize love, how to understand love and how to build it into the machines that we are currently building and that are about to become smarter than us. Well, I think this is a good opportunity to try to sneak up to the idea of enlightenment. So you wrote a series of good tweets about consciousness and panpsychism. So let's break it down. First you say, I suspect the experience that leads to the panpsychism syndrome of some philosophers and other consciousness enthusiasts represents the realization that we don't end at the self but share a resonant universe representation with every other observer coupled to the same universe. This actually eventually leads us to a lot of interesting questions about AI and AGI. But let's start with this representation. What is this resonant universe representation? And what do you think? Do we share such a representation?

The neuroscientist Grossberg has come up with the cognitive architecture that he calls the adaptive resonance theory. And his perspective is that our neurons can be understood as oscillators that are resonating with each other and this outside phenomena. So the coarse-grained model of the universe that we are building in some sense is resonance with objects and outside of us in the world. So basically we take up patterns of the universe that we are coupled with and our brain is not so much understood as circuitry even though this perspective is valid. But it's almost an ether in which the individual neurons are passing on chemical electrical signals or arbitrary signals across all modalities that can be transmitted between cells, simulate each other in this way, and produce patterns that they modulate while passing them on. And this speed of signal progression

in the brain is roughly at the speed of sound, incidentally because the time that it takes for the signals to hop from cell to cell, which means it's relatively slow with respect to the world. It takes an appreciable fraction of a second for a signal to go through the entire neocortex, something like a few hundred milliseconds. And so there's a lot of stuff happening in that time where the signal is passing through your brain, including in the brain itself. So nothing in the brain is assuming that stuff happens simultaneously. Everything in the brain is working in a paradigm

where the world has already moved on when you are ready to do the next thing to your signal, including the signal processing system itself. That's quite a different paradigm

than the one in our digital computers, where we currently assume that your GPU or CPU is pretty much globally in the same state. So you mentioned there the non-dual state and say that some people confuse it for enlightenment. What's the non-dual state?

There is a state in which you notice that you are no longer a person and instead you are one with the universe. So that speaks to the resonance? Yes, but this one with the universe is of course not

accurately modeling that you are indeed some God entity or indeed the universe becoming aware of itself, even though you get this experience. I believe that you get this experience because your mind is modeling the fact that you are no longer identified with the personal self in that state, but you have transcended this division between the self-model and the world-model and you are experiencing yourself as your mind, as something that is representing a universe. But that's still part of the model? Yes. So it's inside of the model still, you are still inside of patterns that are generated in your brain and in your organism, and what you are now experiencing is that you are no longer this personal self in there, but you are the entirety of the mind on its contents. Why is it so hard to get there? A lot of people who get into this state think this or are associated with enlightenment, I suspect it's a favorite training goal for a number of meditators, but I think that enlightenment is in some sense more mundane and it's a step further or sideways. It's the state where you realize that everything is a representation. Yeah, you say enlightenment is a realization of how experience is implemented. Yes, so basically you notice at some point that your qualia can be deconstructed. Reverse engineered, almost like a schematic of it, what? You can start with looking at a face, maybe look at your own face in the mirror, look at your face for a few hours in the mirror or for a few minutes, at some point it will look very weird and because you notice that there's actually no face. You basically start unseeing the face, what you see is a geometry, and then you can disassemble the geometry and realize how that geometry is being constructed in your mind. And you can learn to modify this, so basically you can change these generators in your own mind to shift the face around or to change the construction of the face, to change the vein which the features are being assembled. Why don't we do that more often? Why don't we start really messing with reality without the use of drugs or anything else? Why don't we get good at this kind of thing, like intentionally? Why should we? Because you can morph reality into something more pleasant for yourself, just have fun with it. Yeah, that is probably what you shouldn't be doing, right? Because outside of your personal self, this outer mind is probably a relatively smart agent. And what you often notice is that you have thoughts about how you should live, but you observe yourself doing different things and have different feelings. And that's because your outer mind doesn't believe you and doesn't believe your rational thoughts. Well, can't you just silence the outer mind? The thing is that the outer mind is usually smarter than you are. Rational thinking is very brittle. It's very hard to use logic and symbolic thinking to have an accurate model of the world. So there is often an underlying system that is looking at your rational thoughts and then tells you, no, you're still missing something. Your gut feeling is still facing something else. And this can be, for instance, you find a partner that looks perfect, or you find a deal, and you build a company or whatever that looks perfect to you. And yet at some level, you feel something is off and you cannot put your finger on it. And the more reason about it, the better it looks to you. But the system that is outside still tells you, no, no, you're missing something. And that system is powerful. People call this intuition, right? Intuition is this unreflected part of your attitude, composition and computation, where you produce a model of how you relate to the world and what you need to do it and what you can do in it and what's going to happen, that is usually deeper and often more accurate than your reason. So if we look at this as you write in the tweet, if we look at this more rigorously as a sort of take the panpsychist idea more seriously, almost as a scientific discipline,

you write that quote, fascinatingly, the panpsychist interpretation seems to lead to observations of practical results to a degree that physics fundamentalists might call superstitious. Reports of long distance telepathy and remote causation are ubiquitous in the general population. I am not convinced, says Yoshibak, that establishing the empirical reality of telepathy would force an update of any part of serious academic physics, but it could trigger an important revolution in both neuroscience and AI from a circuit perspective to a coupled, complex resonator paradigm. Are you suggesting that there could be some rigorous mathematical wisdom to panpsychist perspective on the world?
So first of all, panpsychism is the perspective that consciousness is inseparable from matter in the universe. And I find panpsychism quite unsatisfying because it does not explain consciousness, but it does not explain how this aspect of matter produces. It is also when I try to formalize panpsychism and write down what it actually means and with a more formal mathematical language, it's very difficult to distinguish it from saying that there is a software side to the world in the same way as there is software side to what the transistors are doing in your computer. So basically, there is a pattern at a certain core screening of the universe that in some reasons of the universe leads to observers that are observing themselves. So panpsychism maybe is not even when I write it down a position that is distinct from functionalism. But intuitively, a lot of people feel that the activity of matter itself, of mechanisms in the world, is insufficient to explain it. So it's something that needs to be intrinsic to matter itself. And you can, apart from this abstract idea, have an experience in which you experience yourself as being the universe, which I suspect is basically happening because you manage to dissolve
the division between personal self and mind that you establish as an infant when you construct a personal self and transcend it again and understand how it works. But there is something deeper that is that you feel that you're also sharing a state with other people, that you have an experience in which you notice that your personal self is moving into everything else, that you basically look out of the eyes of another person, that every agent in the world that is an observer is in some sense you. And we forget that we are the same agent. So is it that we feel that or do we actually accomplish it? So is telepathy possible? Is it real? So for me, that's the question that I don't really know the answer to. In Turing's famous 1950 paper in which he describes the Turing test, he does speculate about telepathy, interestingly, and asks himself if telepathy is real, and he thinks that it very well might be, but it would be the implication for AI systems that try to be intelligent, because he didn't see a mechanism by which a computer program would become telepathic. And I suspect if telepathy would exist, or if all the reports that you get from people when you ask the normal person on the street, I find that very often they say, I have experiences with telepathy, the scientists might not be interested in this and might not have a theory about this, but I have difficulty explaining it away. And so you could say maybe this is a superstition, maybe it's a false memory, or maybe it's a little bit of psychosis, who knows. Maybe somebody wants to make their own life more interesting or misremember something. But a lot of people report, I noticed something terrible happened to my partner, and I know this is exactly the moment it happened, where my child had an accident, and I knew that was happening and the child was in a different town. So maybe this is a false memory, where this is later on mistakenly attributed, but

a lot of people think that this is not the correct explanation. So if something like this was real, what would it mean? It probably would mean that either your body is an antenna that is sending information over all sorts of channels, like maybe just electromagnetic radio signals that you're sending over long distances, and you get attuned to another person that you spend enough time with to get a few bits out of the ether to figure out what this person is doing. Or maybe it's also when you are very close to somebody and you become empathetic with them, what happens that is that you go into a resonance state with them. Similar to when people go into a seance and they go into a trance state and they start shifting a Ouija board around on the table, I think what happens is that their minds go by their nervous systems into a resonance state in which they basically create something like a shared dream between them.
Physical closeness or closeness broadly defined?
With physical closeness, it's much easier to experience empathy with someone. I suspect it would be difficult for me to have empathy for you if you were in a different town. Also, how would that work? But if you are very close to someone, you pick up all sorts of signals from their body, not just via your eyes, but with your entire body. If the nervous system sits on the other side and the intercellular communication sits on the other side and is integrating over all these signals, you can make inferences about the state of the other. It's not just the personal self that does this via reasoning, but your perceptual system. What basically happens is that your representations are directly interacting. It's the physical resonant models of the universe that exist in your nervous system and in your body might go into resonance with others and start sharing some of their states. Basically, by being next to somebody, you pick up some of their vibes and feel without looking at them what they're feeling in this moment. It's difficult for you if you're very empathetic to detach yourself from it and have an emotional state that is completely independent from your environment. People who are highly empathetic are describing this. Now imagine that a lot of organisms on this planet have representations of the environment and operate like this, and they are adjacent to each other and overlapping. There's going to be some degree in which there is basically some chain interaction and we are forming some slightly shared representation. There are relatively few neuroscientists who consider this possibility.
I think a rarity in this regard is Michael Levin who is considering these things in earnest.
I stumbled on this train of thought mostly by noticing that the tasks of a neuron can be fulfilled by other cells as well. They can send different type chemical messages and physical messages to their adjacent cells and learn when to do this and why not make this conditional and become universal functional approximators. The only thing that they can not do is telegraph information over axons very quickly over long distances. Neurons in this perspective are specially adapted kind of telegraph cell that has evolved so we can move our muscles very fast, but our body is in principle able to also make models of the world just much, much slower.
It's interesting though that at this time at least in human history there seems to be a gap between the tools of science and the subjective experience that people report like you're talking about with telepathy and it seems like we're not quite there. No, I think that there is no gap between the tools of science and telepathy. Either it's there or it's not and it's an empirical question and if it's there we should be able to detect it in a lab. So why is there not a lot of Michael Levin's walking around? I don't think that Michael Levin is specifically focused on telepathy very much. He is focused on self-organization in living organisms and in brains both as a

paradigm for development and as a paradigm for information processing and when you think about how organization processing works in organism there is first of all radical locality which means everything is decided locally from the perspective of an individual cell. The individual cell is the agent and the other one is coherence. Basically there needs to be some criterion that determines how these cells are interacting in such a way that order emerges on the next level of structure and this principle of coherence of imposing constraints that are not validated by the individual parts and lead to coherent structure to basically transcendental agency where you form an agent on the next level of organization is crucial in this perspective.
It's so cool that radical locality leads to the emergence of complexity at the higher layers.
And I think what Michael Levin is looking at is nothing that is outside of the realm of science in any way. It's just that he is a paradigmatic thinker who develops his own paradigm and most of the neuroscientists are using a different paradigm at this point and this often happens in science that a field has a few paradigms in which people try to understand reality and build concepts and make experiments. You're kind of one of those type of paradigmatic thinkers.
Actually if we can take a tangent on that once again returning to the biblical verses of your tweets you're right my public explorations are not driven by audience service but by my lack of ability for discovering understanding or following the relevant authorities. So I have to develop my own thoughts since I think autonomously these thoughts cannot always be very good.
That's you apologizing for the chaos of your thoughts or perhaps not apologizing just identify but let me ask the question. Since we talked about Michael Levin and yourself who I think are very kind of radical big independent thinkers can we reverse engineer your process of thinking autonomously? How do you do it? How can humans do it? How can you avoid being influenced
by what is the stage three? Well why would you want to do that? It's you see what is working for you and if it's not working for you you build another structure that works better for you right and so I found myself and when I was thrown into this world in a state where my intuitions were not working for me I was not able to understand how I would be able to survive in this world and build the things that I was interested in build the kinds of relationship I needed to build work on the topics that I wanted to make progress on and so I had to learn and I for me Twitter is not some tool of publication it's not something where I put stuff that I entirely believe to be true and provable. It's an interactive notebook in which I explore possibilities and I found that when I tried to understand how the mind and how consciousness
works I was quite optimistic I thought there needs to be a big body of knowledge that I can just study and that works and so I entered studies in philosophy and computer science and later psychology and a bit of neuroscience and so on and I was disappointed by what I found because I found that the questions of how consciousness and so on works how emotion works how it's possible that the system can experience anything how motivation emerges in the mind were not being answered by the authorities that I met and the schools that were around and instead I found that it was individual thinkers that had useful ideas that sometimes were good sometimes were not so good sometimes were adopted by a large group of people
sometimes were rejected by large groups of people but for me it was much more interesting to see

these minds as individuals and in my perspective thinking is still something that is done not in groups that has to be done by individuals. So that motivated you to become an individual thinker yourself? I didn't have a choice. I didn't find a group that thought in a way where I felt okay I can just adopt everything that everybody thinks here and now I understand how consciousness
works right so or how the mind works or how thinking works or what thinking even is or what feelings are and how they're implemented and so on. So to figure all this out I had to take a lot of ideas from individuals and then try to put them together and something that works for myself and on one hand I think it helps if you try to go down and find first principles in which you can recreate how thinking works how languages work what representation is whether representation is necessary how the relationship between a representing agent and the world works in general. But how do you escape the influence once again the pressure of the crowd
whether it's you in responding to the the pressure or you being swept up by the pressure if you even just look at twitter the opinions of the crowd. I don't feel pressure from the crowd I'm completely immune to that in the same sense I don't have respect for authority I have respect for what an individual is accomplishing or have respect for mental fire power or so but it's not that I meet somebody and get lectured and unable to speak or when a large group of people has a certain idea that is different for me I don't necessarily feel intimidated which has often been a problem for me in my life because I like instincts that other people develop at a very young age and that help with their self-preservation in a social environment so I had to learn a lot of things the hard way. Yeah so is there a practical advice you can give on how to think paradigmatically how to think independently or you know because you've kind of said I had no choice
but I think to a degree you have a choice
because you said you want to be productive and I think thinking independently is productive if what you're curious about is understanding the world especially when the problems are very kind of new and open so it seems like this is a active process like we can choose to do that we can practice it. Well it's a very basic question when you read a theory that you find convincing or interesting how do you know it's very interesting to figure out what are the sources of that other person not which authority can they refer to that is then taking off the burden of being truthful but how did this authority in turn know what is the epistemic chain to observables what are the first principles from which the whole thing is derived and when I was young I was not blessed with a lot of people around myself who knew how to make proofs from first principles and I think mathematicians
do this quite naturally but most of the great mathematicians do not become mathematicians in school but they tend to be self-taught because school teachers tend not to be mathematicians right they tend not to be people who derive things from first principles so when you ask your school teacher why does two plus two equal four does your school teacher give you the right answer like it's a simple game and many simple games that you could play and most of those games that you could just take different rules would not lead to an interesting arithmetic and so it's just an exploration but you can try what happens if you take different axioms and here is how you build axioms and derive addition from them and a built addition is some basically syntactic sugar

and so this I wish that somebody would have opened me this vista and explained to me how I can build a language in my own mind and from which I can derive what I'm seeing and how I can which I can make geometry and counting and all the number games that we are playing in our life and on the
other hand I felt that I learned a lot of this while I was programming as a child when you start out with a computer like a Commodore 64 which doesn't have a lot of functionality it's relatively easy to see how a bunch of relatively simple circuits are just basically performing hashes between bit patterns and how you can build the entirety of mathematics and computation on top of this and all the representational languages that you need. Commodore 64 could be one of the sexiest machines I've ever built if I so say so myself. If we can return to this really interesting idea that we started to talk about with panpsychism and the complex resonator paradigm and the verses
of your tweets you write instead of treating eyes ears and skin as separate sensory systems with fundamentally different modalities who might understand them as overlapping aspects of the same universe coupled at the same temporal resolution and almost inseparable from a single shared resonant model instead of treating mental representations as fully isolated between minds the representations of physically adjacent observers might directly interact and produce causal effects through the coordination of the perception and behavior of world modeling observers so the modalities the distinction between modalities let's throw that away the distinction between the individuals let's throw that away so what does this interaction representations look like and you think about how you represent the interaction of us in this room yeah at some
level you can the modalities are quite distinct they're not completely distinct but you can see this as vision you can close your eyes and then you don't see a lot anymore but you still imagine how my mouth is moving when you hear something and you know that it's very close to the sound that you can just open your eyes and you get back into this shared merged space and we also have these experiments where we notice that the vein which my lips are moving are affecting how you hear
the sound and also vice versa the sounds that you're hearing have an influence on how you interpret some of the visual features and so these modalities are not separate in your mind they do are merged
at some fundamental level where you are interpreting the entire scene that you are in and your own interactions in the scene are also not completely separate from the interactions of the other individual in the scene but there is some resonance that is going on where we also have a degree of shared mental representations and shared empathy due to being in the same space and having vibes between each other. Vibes so the question though is how deeply
interbind is this multi-modality multi-agent system like how I mean this is going to the telepathy question without the woo-woo meaning of the word telepathy is like how like what's going on here in this room right now. So telepathy would work how could it work yeah right so imagine that all the cells in your body are sending signals in a similar way as neurons are doing right just by touching the other cells and sending chemicals to them the other cells interpreting them learning how to react to them and they learn how to approximate functions in this way and compute behavior for the organisms and this is something that is open to plants as well
and so plants probably have software running on them that is controlling how the plant is working

in a similar way as you have a mind that is controlling how you are behaving in the world
and this spirit of plants which is something that has been very well described by our ancestors
and they found this quite normal but for some reason since the enlightenment we are treating
this notion that there are spirits in nature and the plants have spirits as a superstition
and I think we probably have to rediscover that that plants have software running on them
and we already did right you notice that there is a control system in the plant that
connects every part of the plant to every other part of the plant and produces coherent behavior
in the plant that is of course much much slower than the coherent behavior in an animal like us
that has a nervous system that where everything is synchronized much much faster by the neurons
but what you also notice is that if a plant is sitting next to another plant like you have a
very old tree and this tree is building some kind of information highway along its cells so it can
send information from its leaves to its roots and from some part of the root to another part of the
roots and there is a fungus living next to the tree the fungus can probably piggyback on the
communication between the cells of the tree and send its own signals through the tree
and vice versa the tree might be able to send information to the fungus because after all how
would they build a viable firewall if that other organism is sitting next to them all the time
and is never moving away and so they will have to get along and over a long enough time frame
the networks of roots in the forest and all the plant other plants that are there and the
fungi that are there might be forming something like a biological internet
but the question there is do they have to be touching is biology at a distance possible
of course you can use any kind of physical signal you can use sounds you can use electromagnetic
waves that are integrated over many cells that's conceivable that across distances there are many
kinds of information pathways but also our planetary surface is pretty full of organisms
full of cells so everything is touching everything else and it's been doing this for many millions
and even billions of years so there was enough time for information processing networks to form
and if you think about how a mind is self-organizing basically it needs to in some sense reward the
cells for computing the mind for building the necessary dynamics between the cells that allow
the mind to stabilize itself and remain on there but if you look at the spirits of plants that are
growing very close to each other in the forest that might be almost growing into each other
these spirits might be able even to move to some degree not to become somewhat dislocated
and shift around in that ecosystem right and so if you think about what a mind is it's a
bunch of activation waves that form coherent patterns and process information in a way that
are colonizing an environment well enough to allow the continuous sustenance of the mind the
continuous stability and self-stabilization of the mind then it's conceivable that we can link
into this biological internet not necessarily at the speed of our nervous system but maybe at the
speed of our body and make some kind of subconscious connection to the world where we use our
body as
an antenna into biological information processing now now these ideas are completely speculative
I don't know if any of that is true but if that was true and if you want to explain telepathy I think
it's much more likely that such telepathy could be explained using such mechanisms rather than
undiscovered quantum processes that would break the standard model of physics
could there be undiscovered processes that don't break yeah so if you think about
something like an internet in the forest that is something that is borderline discovered

there are basically a lot of scientists who point out that they do observe that
plants are communicating the forest so wood networks and send information for instance
warn each other about new pests entering the forest and and things are happening like this so
basically there is communication between plants and fungi that has been observed well it's been
observed but we haven't plugged into it so it's like if you observe humans they seem to be
communicating
with a smartphone thing but you don't understand how smartphone works and how the the
mechanism
the internet works bore like maybe it's possible to really understand the full richness of the
biological internet that connects us an interesting question is whether the communication and the
organization principles of biological information processing are as complicated as the technology
that we've built they set up on very different principles right they simultaneously works very
differently in biological systems and the entire thing needs to be stochastic and instead of being
fully deterministic or almost fully deterministic as our digital computers are so there is a different
base protocol layer that would emerge over the biological structure if such a thing would be
happening and again i'm not saying here that telepathy works and not saying that this is not
but what i'm saying is i'm i think i'm open to a possibility that we see that a few bits can be
traveling long distance between organisms using biological information processing in ways
that we are not completely aware of right now and that are more similar to many of the stories
that were completely normal for our ancestors well this kind of interacting,
intertwined representations takes us to the the big ending of your tweet series you write quote
i wonder if self-improving agi might end up saturating physical environments with intelligence
to such a degree that isolation of individual mental states becomes almost impossible
and the representations of all complex self-organizing agents merge permanently with each other
so that's a really interesting idea this biological network
life network gets so dense that it might as well be seen as one
that's an interesting uh what do you think that looks like what do you think that saturation
looks like what does it feel like i think it's a possibility it's just a vague possibility
and i like to explain but um what this looks like i think that the end game of agi is
substrate agnostic that means that agi ultimately if it is being built is going to be smart enough
to understand how agi works this means it's not going to be better than people at agi research
and can take over and building the next generation but it fully understands how it works and how it's
being implemented and also of course understands how computation works in nature how to build
new
feedback loops that you can turn into your own circuits and this means that the agi is likely
to virtualize itself into any environment that can compute so it's not breaking free
from the silicon substrate and is going to move into the ecosystems into our bodies our brains
and is going to merge with all the agency that it finds there yeah so it's conceivable that you
end up with completely integrated information processing across all computing systems including
biological computation on earth you that we end up triggering some new step in the evolution where
basically some Gaia is being built over the entirety of all digital and biological computation
and if this happens then basically everywhere around us you will have agents that are connected
and that are representing and building models of the world and their representations will physically

interact they will vibe with each other and if you find yourself into an environment in an environment that is saturated with modeling compute where basically you almost every grain of sand could be part of computation that is at some point being started by the ai you could find yourself in a situation where you cannot escape this shared representation anymore and where you indeed notice that everything in the world has one shared resonant model of everything that's happening on the planet and you notice which part you are in this thing and you become part of a very large or almost holographic mind in which all the parts are observing each other and form a coherent whole so you lose the ability to notice to notice yourself as a distinct entity
no i think that when you are conscious in your own mind you notice yourself as a distinct entity you notice yourself as a self-reflexive observer and i suspect that we become conscious at the beginning of our mental development not at some very high level consciousness seems to be part of a training mechanism that biological nervous systems have to discover to become trainable because you cannot take a nervous system like ours and to stochastic rate at the center spec propagation over 100 layers and this would not be stable on biological neurons and so instead we start with some colonizing principle in which a part of the mental representations form a notion of being a self-reflexive observer that is imposing coherence on its environment and this spreads until the boundary of your mind and if that boundary is no longer clear cut because ai is jumping across substrates it would be interesting to see what a global mind would look like that is basically producing a globally coherent language of thought and is representing everything from all the possible
vantage points. That's an interesting world. The intuition that this thing go out of is a particular mental state and it's a state that you find sometimes in literature for instance Neil Gaiman describes it in the ocean at the end of the lane and it's this idea that or this experience that there is a state in which you feel that you know everything that can be known and that in your normal human mind you've only forgotten that you are the entire universe
and some people describe this after they've taken extremely large amount of mushrooms or had a big
spiritual experience as hippie in their 20s and they notice basically that they are in everything and their body is only one part of the universe and nothing ends at their body and actually everything is observing and they are part of this big observer and the big observer is focused as one local point in their body and their personality and so on but we can basically have this oceanic state in which you have no boundaries and are one with everything and a lot of meditators call this the non-dual state because you no longer have the separation between self and world and as I said you can explain the state relatively simply without panpsychism or anything else but just by breaking down the constructed boundary between self and world in our own mind but if you combine this with the notion that the systems are physically
interacting to the point where their representations are merging and interacting with each other you would literally implement something like this. It would still be a representational state where you would not be one with physics itself, it would still be coarse-grained, it would still be much slower than physics itself but it would be a representation in which you become aware that you're part of some kind of global information processing system like thought in a global mind and a conscious thought that co-existing with many other self-reflexive thoughts.

Just I would love to observe that from a video game design perspective how that game looks.
Maybe you will after we build AGI and it takes over.
But would you be able to step away, step out at the whole thing, just kind of watch,
you know the way we can now sometimes when I'm in a crowded party or something like this
you step back and you realize all the different costumes, all the different interactions,
all the different computation that all the individual people are once distinct from each
other and at once all the same. But it's already what we do, right? We can have thoughts that are integrative and we have kind of thoughts that are highly dissociated from everything else and experience themselves as separate. But you want to allow yourself to have those thoughts.
Sometimes you kind of resist it. I think that it's not normative. It's more descriptive.
I want to understand the space of states that we can be in and that people are reporting
and make sense of them. It's not that I believe that it's your job in life to get to a particular
kind of state and then you get a high score. Or maybe you do. I think you're really
against this high scoring thing. Yeah, you're probably very competitive and I'm not.
No, not competitive. Like role-playing games like Skyrim, it's not competitive. There's a
nice feeling where your experience points go up. You're not competing against anybody,
but it's the world saying you're on the right track. Here's a point.
That's the game saying it. It's the game economy. And I found when I was playing games and was getting addicted to these systems, then I would get into the game and hack it. So I get control over the scoring system and would no longer be subject to it. So you're no longer playing, you're trying to hack it. I don't want to be addicted to anything. I want to be in charge. I want to have agency over what I do. Addiction is the loss of control for you. Yes, addiction means that you're doing something compulsively. And the opposite of free will is not determinism. It's compulsion.
You don't want to lose yourself in the addiction to something nice,
addiction to love, to the pleasant feelings we humans experience.
No, I find this gets old. I don't want to have the best possible emotions. I want to have the most appropriate emotions. I don't want to have the best possible experience. I want to have an adequate experience that is serving my goals, the stuff that I find meaningful in this world.
From the biggest questions of consciousness, let's explore the pragmatic, the projections of those big ideas into our current world. What do you think about LLMs, the recent rapid development of large language models, of the AI world, of generative AI? How much of the hype is deserved and how much is not? And people should definitely follow your Twitter because you explore these questions in a beautiful, profound, and hilarious way at times.
No, don't follow my Twitter. I already have too many followers. At some point,
it's going to be unpleasant. I noticed that a lot of people feel that it's totally okay to punch up and it's a very weird notion that you feel that you haven't changed, but your account has grown and suddenly you have a lot of people who casually abuse you. I don't like that I have to block more than before and I don't like this overall vibe shift. And right now, it's still somewhat okay, so pretty much okay, so I can go to a place where people work on stuff that I'm interested in and as a good chance that a few people in the room know me so there's no awkwardness. But when I get to a point where random strangers feel that they have to have an opinion about me one way or the other, I don't think I would like that. And random strangers, because of your kind of out in their mind elevated position? Yes, so basically whenever you are in any way prominent or some

kind of celebrity, random strangers will have to have an opinion about you.
Yeah, and they kind of forget that you're human too. I mean, you notice this thing yourself that the more popular you get, the higher the pressure becomes, the more winds are blowing in your direction from all sides. And it's stressful, right? And it does have a little bit of upside, but it also has a lot of downside. I think it has a lot of upside, at least for me currently, at least perhaps because of the podcast, because most people are really good and people come up to me and they have love in their eyes and over a stretch of like 30 seconds, you can hug it out and you can just exchange a few words and you reinvigorate your love for humanity.
So that's an upside for a loner. I'm gonna look because otherwise you have to do a lot of work to find such humans. And here you're like thrust into the full humanity, the goodness of humanity for the most part. Of course, maybe guess worse as you become more prominent. I hope not.
This is pretty awesome. I have a couple handful very close friends and I don't have enough time for them and attention for them as it is. And I find this very, very regrettable. And then there are so many awesome, interesting people that I keep meeting. And I would like to integrate them in my life, but I just don't know how because there's only so much time and attention. And the older I get, the harder is to bond with new people in a deep way. But can you enjoy, I mean, there's a picture of you, I think with Roger Penrose and Eric Weinstein and a few others that are interesting figures. Can't you just enjoy random, interesting humans for a short amount of time? I'm also, I like these people and what I like is intellectual stimulation. And I'm very grateful that I'm getting it. Can you not be melancholy or maybe I'm projecting, I hate goodbyes. Can we just not hate goodbyes and just enjoy the hello, take it in, taking a person, taking their ideas and then move on through life? I think it's totally okay to be sad about goodbyes because that indicates that there was something that you're going to miss.
Yeah, but it's painful. Maybe that's one of the reasons I'm an introvert is I hate goodbyes.
But you have to say goodbye before you say hello again.
I know. But at that experience of loss, that many loss, maybe that's a little death. Maybe, I don't know, I think this melancholy feeling is just the other side of love. And I think they go hand in hand and it's a beautiful thing. And I'm just being romantic about it at the moment. And I'm not no stranger to melancholy. And sometimes, it's difficult to bear to be alive. Sometimes it's just painful to exist.
But there's beauty in that pain too. That's what melancholy feeling is. It's not negative.
Like melancholy doesn't have to be negative. Can also kill you. Well, we all die eventually.
Now, as we got to this topic, the actual question was about what your thoughts are about the development, the recent development of large language models with chat GPT. There's a lot of hype. Is some of the hype justified? Which is, which isn't? What are your thoughts? High level? I find that large language models do have this coding, right? So it's an extremely useful application that is for a lot of people taking stack overflow out of their life and exchange for something that is more efficient. I feel that chat GPT is like an intern that I have to micromanage. I have been working with people in the past who were less capable than chat GPT. And I'm not saying this because I hate people, but they personally as human beings, there was something present that was not there in chat GPT, which was why I was covering for them.
But chat GPT is, has an interesting ability. It does give people superpowers. And the people

who feel threatened by them are the prompt completers. They are the people who do what chat GPT is doing right now. So if you are not creative, if you don't build your own thoughts, if you don't have actual plans in the world, and your only job is to summarize emails and to expand simple intentions into emails again, then chat GPT might look like a threat. But I believe that it is a very beneficial technology that allows us to create more interesting stuff and make the world more beautiful and fascinating if we find to build it into our life in the right ways. So I'm quite fascinated by these large language models, but I also think that they are by no means the final development. And it's interesting to see how this development progresses. One thing that the out of the box vanilla language models have as a limitation is that they have still some limited coherence and ability to construct complexity. And even though they exceed human abilities to do what they can do one shot. Typically, when you write a text with a language model or using it or when you write code for the language model, it's not one shot because they're going to be bugs in your program and design errors and compiler errors and so on. And your language model can help you to fix those things. But this process is out of the box not automated yet. So there is a management process that also needs to be done. And there are some interesting developments, baby AGI and so on, that are trying to automate this management process as well. And I suspect that soon we are going to see a bunch of cognitive architectures where every module is in some sense a language model or something equivalent. And between the language models, we exchange suitable data structures, not English and produce compound behavior of this whole thing. To do some of the quote unquote prompt engineering for you. They create these kinds of cognitive architectures that do the prompt engineering and you're just doing the high, high level meta prompt engineering. There are limitations in a language model alone.

I feel that part of my mind works similarly to a language model, which means I can yell into it a prompt and it's going to give me a creative response. But I have to do something with those points first. I have to take it as a generative artifact that may or may not be true. It's usually a confibulation. It's just an idea. And then I take this idea and modify it. I might build a new prompt that is stepping off this idea and develops it to the next level or it put it into something larger or I might try to prove whether it's true or make an experiment. And this is what the language models right now are not doing yet. But there's also no technical reason for why they shouldn't be able to do this. So the way to make a language model coherent is probably not to use reinforcement learning until it only gives you one possible answer that is linking to its source data. But it's using this as a component in a larger system that can also be built by the language model or is enabled by a language model structured components or using different technologies. I suspect that language models will be an important stepping stone in developing different types of systems. And one thing that is really missing in the form of language models that we have today is real time world coupling. It's difficult to do perception with a language model and motor control with a language model. Instead, you would need to have different type of thing that is working with it. Also the language model is a little bit obscuring what its actual functionality is. Some people associate the structure of the neural network of the language model with the nervous system and I think that's the wrong intuition. The neural networks are unlike nervous system. They are more like hundred step functions that use differentiable linear algebra to approximate

correlation between adjacent brain states. It's basically a function that moves the step system from one representational state to the next representational state. If you try to map this into a metaphor that is closer to our brain, imagine that you would take a language model or a model like Dali that you use for instance this image guided diffusion to approximate a camera image and use the activation state of the neural network to interpret the camera image which in principle I think will be possible very soon. You do this periodically and now you look at these patterns how when this thing interacts with the world periodically look like it's in time and these time slices they are somewhat equivalent to the activation state of the brain at a given moment.

How's the actual brain different? Just the asynchronous craziness?

For me it's fascinating that they are so vastly different and yet in some circumstances produce somewhat similar behavior and the brain is first of all different because it's a self organizing system where the individual cell is an agent that is communicating with the other agents around it and is always trying to find some solution and all the structure that pops up is emergent structure. One way in which you could try to look at this is that individual neurons probably need to get a reward so they become trainable which means they have to have inputs that are not affecting the metabolism of the cell directly but there are messages, semantic messages that tell the cell whether it's done good or bad and in which direction it should shift its behavior. Once you have such an input neurons become trainable and you can train them to perform computations by exchanging messages with other neurons and parts of the signals that they are exchanging and parts of the computation that are performing are control messages that perform management tasks for other neurons and other cells. I also suspect that the brain does not stop at the boundary of neurons to other cells but many adjacent cells will be involved intermittently in the functionality of the brain and will be instrumental in distributing rewards and in managing its functionality. It's fascinating to think about what those characters of the brain enable you to do that language models cannot do.

So first of all there's a different loss function at work when we learn and to me it's fascinating that you can build a system that looks at 800 million pictures and captions and correlates them because I don't think that a human nervous system could do this. For us the world is only learnable because the adjacent frames are related and we can afford to discard most of that information

during learning. We basically take only in stuff that makes us more coherent not less coherent and our neural networks are willing to look at data that is not making the neural network coherent at first but only in the long run. By doing lots and lots of statistics eventually patterns become visible and emerge and our mind seems to be focused on finding the patterns as early as possible. Yeah so filtering early on not later. Yes it's a slightly different paradigm and it leads to much faster convergence so we only need to look at the tiny fraction of the data to become coherent and of course we do not have the same richness as our train models. We will not incorporate the entirety of text in the internet and be able to refer to it and have all this knowledge available and being able to confirm relate over it. Instead we have a much much smaller part of it that is more deliberately built and to me it would be fascinating to think about how to build such systems. It's not obvious that they would necessarily be more efficient than us on a digital substrate but I suspect that they might so I suspect that the actual AGI that is

going to be more interesting is going to use slightly different algorithmic paradigms or sometimes massively different algorithmic paradigms than the current generation of transformer based learning systems. Do you think it might be using just a bunch of language models like this? Do you think the current transformer based large language models will take us to AGI? My main issue is I think that they're quite ugly and brutalist. Which brutalists that we said? Yes they are basically boot forcing the problem of thought and by training this thing with looking at instances where people have thought and then trying to deep fake that and if you have enough data the deep fake becomes indistinguishable from the actual phenomenon and in many circumstances it's going to be identical. Can you deep fake it till you make it?
So can you achieve what are the limitations of this? I mean can you reason? Let's use words that are loaded. Yes that's a very interesting question. I think that these models are clearly making some inference but if you give them a reasoning task it's often difficult for the experimenters to figure out whether the reasoning is the result of the emulation of the reasoning strategy that this saw in human written text or whether it's something that the system was able to infer by itself. On the other hand if you think of human reasoning, if you want to become a very good reasoner you don't do this by just figuring out yourself. You read about reasoning and the first people who tried to write about reasoning and reflect on it didn't get it right. Even Aristotle who thought about this very hard and came up with the theory of how syllogisms works and syllogistic reasoning has mistakes in his attempt to build something like a formal logic and gets maybe 80% right and the people that are talking about reasoning professionally today read Tarski and Frager and built on their work. So in many ways people when they perform reasoning are emulating what other people wrote about reasoning. So it's difficult to really draw this boundary. And when François Chollet says that these models are only interpolating between what they saw and what other people are doing, well if you give them all the latent dimensions that can be extracted from the internet, what's missing? Maybe there is almost everything there and if you're not sufficiently informed by these dimensions and you need more, I think that's not difficult to increase the temperature in the large-angle model to the point that is producing stuff that is maybe 90% nonsense and 10% viable and combine this with some poover that is trying to
filter out the viable parts from the nonsense in the same way as our own thinking works, right? When
we're very creative we increase the temperature in our own mind and recreate hypothetical universes
and solutions most of which will not work. And then we test and we test by building a core that is internally coherent and we use reasoning strategies that use some axomatic consistency by which we can identify those strategies and thoughts and subuniverses that are viable and that can expand our thinking. So if you look at the language models they have clear limitations right now. One of them is they're not coupled to the world in real time in the way in which our nervous systems are. So it's difficult for them to observe themselves in the universe and to observe what kind of universe they're in. Second, they don't do real-time learnings. They basically get only trained with algorithms that rely on the data being available in batches so it can be parallelized and run sufficiently on the network and so on and real-time learning would be very slow so far and inefficient. That's clearly something that our nervous systems can do to some degree.

And there is a problem with these models being coherent and I suspect that all these problems are solvable without a technological revolution. We don't need fundamentally new algorithms to change that. For instance, you can enlarge the context window and thereby basically create working memory in which you train everything that happens during the day. And if that is not sufficient you add a database and you write some clever mechanisms that the system learns to use to swap out in and out stuff from its prompt context. And if that is not sufficient, if your database is full in the evening, overnight you just train. The system is going to sleep and dream and it's going to train the stuff from its database into the louder model by fine-tuning it, building additional layers and so on. And then the next day it starts with a fresh database and the morning with fresh eyes has integrated all this stuff. And when you talk to people and you have strong disagreements about something which means that in their mind they have a faulty belief or you have a faulty belief with a lot of dependencies on it, very often you will not achieve agreement in one session, but you need to sleep about this once or multiple times before you have integrated all these necessary changes in your mind. So maybe it's already somewhat similar. Yeah, there's already a latency even for humans to update the model. We train the model. And of course we can combine the language model with models that get coupled to reality in real time and can build multi-modal model and bridge between vision models and language models and so
on. So there is no reason to believe that the language models will necessarily run into some problem that will prevent them from becoming generally intelligent. But I don't know that.
It's just I don't see proof that there wouldn't. My issue is I don't like them. I think that they're inefficient. I think that they use way too much compute. I think that given the amazing hardware that we have, we could build something that is much more beautiful than our own mind and
this thing is not as beautiful as our own mind, despite being so much larger.
But it's a kind of proof of concept. It's the only thing that works right now. So it's not the only game in town, but it's the only thing that has this utility with so much simplicity. There's a bunch of relatively simple algorithms that you can understand in relatively few weeks that can be scaled up massively. So it's the deep blue of chess playing. Yeah, it's ugly.
Yeah, Claude Shannon had this when he described chess suggested that there are two main strategies
in which you could play chess. One is that you are making a very complicated plan that reaches far into the future and you try not to make a mistake while enacting it. And this is basically the human strategy. And the other strategy is that you are brute forcing your way to success, which means you make a tree of possible moves where you look at in principle every move that is open to you
or the possible answers. And you try to make this as deeply as possible. Of course, you optimize, you cut off trees that don't look very promising and use libraries of end game and early game and so on to optimize this entire process. But this brute force strategy is how most of the chess programs were built. And this is how computers get better than humans at playing chess. And I look at the large language models, I feel that I'm observing the same thing. It's basically the brute force strategy to sort by training the thing on pretty much the entire internet. And then in the limit, it gets coherent to a degree that approaches human coherence. And on a side effect,

it's able to do things that no human could do. It's able to sift through massive amounts of text relatively quickly and summarize them quickly. And it's never lapses in attention. And I still have the illusion that when I play with chat GPT that it's in principle not doing anything that I could not do if I had Google at my disposal and I get all the resources from the internet and spend enough time on it. But this thing that I have an extremely autistic, stupid intern in a way that is extremely good at drudgery. And I can upload the drudgery to the degree that I'm able to automate the management of the intern is something that is difficult for me to overhype at this point because we have not yet started to scratch the surface of what's possible with this. But it feels like it's a tireless intern or maybe it's an army of interns. And so you get to command these slightly incompetent creatures. And there's an aspect because of how rapidly you can iterate with it. It's also part of the brainstorming part of the kind of inspiration for your own thinking. So you get to interact with the thing. I mean, when I'm programming or doing any kind of generational GPT is it's somehow is a catalyst for your own thinking in the way that I think an intern might not be. Yeah, it gets really interesting. I find is when you turn it into a margin agent system. So for instance, you can get the system to generate a dialogue between a patient and a doctor very easily. But what's more interesting is you have one instance of chat GPT that is a patient and you tell it in the prompt what kind of complicated syndrome it has. And the other one is a therapist who doesn't know anything about this patient. And you just have these two instances battling it out and observe the psychiatrist or psychologist trying to analyze the patient and trying to figure out what's wrong with the patient. And if you try to take a very large problem, a problem, for instance, how to build a company and you turn this into lots and lots of sub problems, then often you can get to a level where the language model is able to solve this. What I also found interesting is based on the observation that chat GPT is pretty good at translating between programming languages. But sometimes it's difficult to write very long coherent algorithms that you need to write them as human author. Why not design a language that is suitable for this? So some kind of pseudocode that is more relaxed than Python. And that allows you to sometimes specify a problem vaguely in human terms and let the chat GPT take care of the rest. And you can use chat GPT to develop that syntax for it and develop new kinds of programming paradigms in this way. So we very soon get to the point where this question, the age old question for us computer scientists, what is the best programming language and can we write a better programming language now that this, I think that almost every serious computer scientist goes to a phase like this in their life. This is a question that is almost no longer relevant. Because what is different between the programming languages is not what they let the computer do, but what they let you think about what the computer should be doing. And now the chat GPT becomes an interface to this in which you can specify in many, many ways what the computer should be doing and chat GPT or some other language model or combination of system is going to take care of the rest. And allow you to expand the realm of thought you're allowed to have when interacting with the computer. It sounds to me like you're saying there's basically no limitations your intuition says to what large language I don't know of that limitation. So when I currently play with it, it's quite limited. I wish that it was way better. But isn't that your fault versus the larger? No, of course, it's always my fault. There's probably a way to make everything. I just want to get you on the record. Yes, everything is my fault. That works doesn't work in my life.

At least that is usually the most useful perspective for myself. Even though the sign site I feel no, I sometimes wish I could have seen myself as part of my environment more and understand that a lot of people are actually seeing me and looking at me and not trying to make my life work in the same way as I try to help others. And making this switch to this levels three perspective is something that happened long after my level four perspective in my life. And I wish that I could have had it earlier. And it's also not now that I don't feel like I'm complete. I'm all over the place. That's all. Worst happiness in terms of stages is on three and four. No, you can be happy at any stage or unhappy. But I think that if you are at a stage where you get agency over how your feelings are generated, and to some degree, you start doing this when you leave adolescence, I believe, that you understand that you're in charge of your own emotion to some degree, and that you are responsible how you approach the world, that it's basically your task to have some basic hygiene, how in the way in which you deal with your mind, and you cannot blame your environment for the way in which you feel. But you live in a world that is highly mobile, and it's your job to choose the environment that you thrive and to build it. And sometimes it's difficult to get the necessary strength and energy to do this, and independence and the worst you feel the harder it is. But it's something that we learn.
It's also this thing that we are usually incomplete. I'm a rare mind, which means I'm mind that is incomplete in ways that are harder to complete. So for me, it might have been harder to initially to find the right relationships and friends that complete me to the degree that I become an almost functional human being. Oh, man, the search space of humans that complete you is an interesting one, especially for Yoshibak. That's an interesting, because talking about brute force search in chess, I wonder what that search tree looks like.
I think that my rational thinking is not good enough to solve that task. A lot of problems in my life that I can conceptualize as software problems, and the failure modes are bugs, and I can debug them and write software that take care of the missing functionality. But there is stuff that I don't understand well enough to use my analytical reasoning to solve the issue. And then I have to develop my intuitions, and often I have to do this with people who are wiser than me. And that's something that's hard for me, because I'm not born with the instinct to submit to other people's wisdom. Yeah. So what kind of problems are we talking about? This is stage three, like love? I found love is never hard.
What is hard then?
Fitting into a world where most people work differently than you and have different intuitions of what should be done. So empathy. It's also aesthetics. When you come into a world where almost everything is ugly, and you come out of a world where everything is beautiful, I grew up in a beautiful place as a child of an artist. And in this place, it was mostly nature. Everything had intrinsic beauty. And everything was built out of an intrinsic need for it to work for itself. Everything that my father created was something that he made to get the world to work for himself. And I felt the same thing. And when I come out into the world, and I am asked to submit to lots and lots of rules, I'm asking, okay, when I observe your stupid rules, what is the benefit? And I see the life that is being offered as a reward. It's not attractive. When you were born and raised in extraterrestrial prints, in a world full of people wearing suits. So it's a challenging integration.
But it also means that I'm often blind for the ways in which everybody is creating their own

bubble of wholesomeness, or almost everybody, and people are trying to do it. And for me to discover this, it was necessary that I found people who had a similar shape of soul as myself. These are my people, people that treat each other in such a way as if they're around for each other for eternity. How long does it take you to detect the geometry, the shape of the soul of another human, to notice that they might be one of your kind? Sometimes it's instantly and I'm wrong, and sometimes it takes a long time. You believe in love at first sight, Niosha Bach?
Yes. But I also noticed that I have been wrong. So sometimes I look at a person and I'm just enamored by everything about them. And sometimes this persists, and sometimes it doesn't. And I have the illusion that they're much better at recognizing who people are as they grow older.
But that could be just cynicism? No. No, it's not cynicism. It's often more that I'm able to recognize what somebody needs when we interact and how we can meaningfully interact. That's not cynical at all. You're better at noticing. Yes. I'm much better, I think, in some circumstances at understanding how to interact with other people than I did when I was young.
Doesn't mean that I'm always very good at it. So that takes us back to prompt engineering of noticing how to be a better prompt engineer of an LLM. A sense I have is that there's a bottomless well of skill to become a great prompt engineer. It feels like it is all my fault whenever I fail to use JGPT correctly, that I didn't find the right words.
Most of the stuff that I'm doing in my life doesn't need JGPT. There are a few tasks that are where it helps. But the main stuff that I need to do, like developing my own thoughts and aesthetics and relationship to people, it's necessary for me to write for myself. Because writing is not so much about producing an artifact that other people can use, but it's a way to structure your own thoughts and develop yourself. So I think this idea that kids are writing their own essays with JGPT in the future is going to have this drawback that they miss out on the ability to structure their own minds via writing. I hope that the schools that our kids are in will retain the wisdom of understanding what parts should be automated and which ones shouldn't. But at the same time, it feels like there's power in disagreeing with the thing that JGPT produces. So I use it like that for programming. I'll see the thing it recommends and then I'll write different code that disagree. And in the disagreement, your mind grows stronger.
I recently wrote a tool that is using the camera on my MacBook and Swift to read pixels out of it and manipulate them and so on. And I don't know Swift. So it was super helpful to have this thing that is writing stuff for me. And also interesting that mostly it didn't work at first. I felt like I was talking to a human being who was trying to hack this on my computer without understanding my
configuration very much and also making a lot of mistakes. And sometimes it's a little bit incoherent. So you have to ultimately understand what it's doing. That's still no other way around it. But I do feel it's much more powerful and faster than using Stack Overflow.
Do you think GPTN can achieve consciousness?
GPTN probably it's not even clear for the present systems. When I talk to my friends at OpenAI, they feel that this question whether the models currently are conscious is much more complicated than many people might think. I guess that it's not that OpenAI has the homogeneous opinion about this. But there are some aspects to this. One is, of course, this language model has written a lot of text in which people were conscious or describe their own consciousness and it's emulating this. And if it's conscious, it's probably not conscious in a way that is close to the way in which human

beings are conscious. But while it is going through these states and going through 100 step function that is emulating adjacent brain states that require a degree of self-reflection, it can also create a model of an observer that is reflecting itself in real time and describe what that's like. And while this model is a deep fake, our own consciousness is also as if it's virtual, right? It's not physical. Our consciousness is a representation of a self-reflective observer that only exists in patterns of interaction between cells. So it is not a physical object in a sense that exists in base reality, but it's really a representational object that develops its causal power only from a certain modeling perspective. To which degree is the virtuality of the consciousness in chat GPT more virtual and less causal than the virtuality of our own consciousness. But you could say it doesn't count. It doesn't count much more than the consciousness of a character in a novel, right? It's important for the reader to have the outcome, the artifact of a model is described in the text generated by the author of the book, what it's like to be conscious in a particular situation and performs the necessary inferences. But the task of creating coherence in real time in a self-organizing system by keeping yourself coherent, so the system is reflexive, that is something that language models don't need to do. So there is no causal need for the system to be conscious in the same way as we are. And for me it would be very interesting to experiment with this to basically build a system like a cat, probably should be careful at first, build something that's small, that's limited, that's limited resources that we can control and study how systems notice a self-model, how they become self-aware in real time. And I think it might be a good idea to not start with a language model, but to start from scratch using principles of self-organization. Is it okay? Can you elaborate why you think that it's a self-organization? So this kind of radical locality that you see in the biological systems, why can't you start with a language model? What's your intuition? My intuition is that the language models that we are building are golems. They are machines that you give a task and they're going to execute the task until some condition is met and there's nobody home. And the way in which nobody is home leads to that system doing things that are undesirable in a particular context. So you have that thing talking to a child and maybe it says something that could be shocking and traumatic to the child, or you have that thing writing a speech and it introduces errors in the speech that human being would ever do if they were responsible. But the system doesn't know who's talking to whom. There is no ground truth that the system is embedded into. And of course, we can create an external tool that is prompting our language model always into the same semblance of ground truth. But it's not like the internal structure is causally produced by the needs of a being to survive in the universe. It is produced by imitating structure on the internet.
Yeah, but so can we externally inject into it this kind of coherent approximation of a world model that has to sync up? Maybe it is sufficient to use the transformer with a different dose function that optimizes for short-term coherence rather than next token prediction over the long run. We had many definitions of intelligence and history of AI. Next token prediction was not very high up on them. And there are some similarities like a condition as data compression is an old trope, Solomonov induction, where you are trying to understand intelligence as predicting future observations from past observations, which is intrinsic to data compression. And predictive coding is a paradigm with the boundary between neuroscience and physics and computer science. So it's not something that is completely alien. But this radical thing that you only do

next token prediction and see what happens is something where most people, I think, were surprised that this works so well. So simple. But is it really that much more radical than just the idea of intelligence as compression? The idea that compression is sufficient to produce all the desired behaviors is a very radical idea. But equally radical as the next token prediction? It's something that wouldn't work in biological organisms, I believe.
Biological organisms have something like next frame prediction for our perceptual system, where we try to filter out principal components out of the perceptual data and build hierarchies over them to track the world. But our behavior ultimately is directed by hundreds of physiological and probably dozens of social and a few cognitive needs that are intrinsic to us, that are built into the system as reflexes and direct us until we can transcend them and replace them by instrumental behavior that relates to our higher goals. And it also seems so much more complicated and messy than next frame prediction, even the idea of frame seems counter biological. Yes, of course, there's not this degree of simultaneity in the biological system. But again, I don't know whether this is actually an optimization if we imitate biology here, because creating something like simultaneity is necessary for many processes that happen in the brain. And you see the outcome of that by synchronized brain waves, which suggests that there is indeed synchronization going on, but the synchronization creates overhead and this overhead is going to make the cells more expensive to run, and you need more redundancy and it makes the system slower. So if you can build a system in which the simultaneous knee gets engineered into it, maybe you have a benefit that you can exploit that is not available to the biological system and that you should not discard right away. You tweeted once again, quote, when I talked to Chad GPT, I'm talking to an NPC. What's going to be interesting, and perhaps scary, is when AI becomes a first person player. So what does that step look like? I'd really like that tweet.
That step between NPC to first person player. What's required for that? Is that kind of what we've been talking about? This kind of external source of coherence and inspiration of how to take the leap into the unknown that we humans do, the search, man's search for meaning, LLM's search for meaning. I don't know if the language model is the right paradigm because it is doing too much, it's giving you too much, and it's hard once you have too much to take away from it again. The way in which our own mind works is not that we train a language model in our own mind, and after the language model is there, we build a personal self on top of it that then relates to the world. There is something that is being built. There is a game engine that is being built, there is a language of thought that is being developed that allows different parts of the mind to talk to each other, and this is a bit of a speculative hypothesis that this language of thought is there, but I suspect that it's important for the way in which our own minds work. Building these principles into a system might be a more straightforward way to a first-person AI, to something that first creates an attentional self and then creates a personal self.
The way in which this seems to be working is that when the game engine is built in your mind, it's not just following radiance where you are stimulated by the environment and then end up with having a solution to how the world works. I suspect that building this game engine in your own mind does require intelligence. It's a constructive task where at times you need to reason. This is a task that we are fulfilling in the first years of our life.
During the first year of its life, an infant is building a lot of structure about the world that

does inquire experiments and some first principles reasoning and so on. In this time, there is usually no personal self. There is a first-person perspective, but it's not a person. This notion that you are a human being that is interacting in a social context and is confronted with an immutable world in which objects are fixed and can no longer be changed, in which the dream can no longer be influenced is something that emerges a little bit later in our life. I personally suspect that this is something that our ancestors had known and we have forgotten because I suspect that it's there in plain sight in Genesis 1 in this first book of the Bible where it's being described that this creative spirit is hovering over the substrate and then is creating a boundary between the world model and sphere of ideas, earth and heaven as they're being described there and then it's creating contrast and then dimensions and then space and then it creates organic shapes and solids and liquids and builds a world from them and creates plants and animals, gives them all their names. Once that's done, it creates another spirit in its own image, but it creates it as man and woman as something that thinks of itself as a human being and puts it into this world. And the Christians mistranslate this, I suspect, when they say this is the description of the creation of the physical universe by a supernatural being. I think this is literally a description of how in every mind a universe is being created as some kind of game engine by a creative spirit, our first consciousness that emerges in our mind even before we are born and that creates the interaction between organism and world and once that is built and trained, the personal self is being created and we only remember being the personal self. We no longer remember how we created the game engine. So God in this view is the first creative mind in the early days, in the early months of development. And it's still there. You still have this outer mind that creates a sense of whether you're being loved by the world or not and what your place in the world is. It's something that is not yourself that is producing this. It's your mind that does it. So there is an outer mind that basically is an agent that determines who you are with respect to the world and while you are stuck being that personal self in this world until you get to stage six and to destroy the boundary. And we all do this, I think, earlier in small glimpses and maybe sometimes we can remember what it was like when we were a small child and get some glimpses into how it's been. But for most people, that rarely happens.

Just glimpses. You tweeted, quote, suffering results for one part of the mind, failing at regulating another part of the mind. Suffering happens in an early stage of mental development. I don't think that superhuman AI would suffer. What's your intuition there? The philosopher Thomas Metzinger is very concerned that the creation of superhuman intelligence would lead to superhuman suffering. And so he's strongly against it. And personally, I don't think that this happens because suffering is not happening at the boundary between our self and the physical universe. It's not stuff on our skin that makes us suffer. It happens at the boundary between self and world. And the world here is the world model. It's the stuff that is created by your mind. The representation of how the universe is and how it should be and how you yourself relate to this. And at this boundary is where suffering happens. So suffering in some sense is self-inflicted, but not by your personal self. It's inflicted by the mind on the personal self that experiences itself as you. And you can turn off suffering when you are able to get on this outer level. So when you

manage to understand how the mind is producing pain and pleasure and fear and love and so on, then you can take charge of this and you get agency over there yourself.
Technically, what pain and pleasure is, they are learning signals, right? The part of your brain is sending a learning signal to another part of the brain to improve its performance. And sometimes this doesn't work because this trainer who sends the signal does not have a good model of how to improve the performance. So it's sending a signal, but the performance doesn't
get better. And then it might crank up the pain and it gets worse and worse. And the behavior of the system may be even deteriorating as a result. But until this is resolved, this regulation issue, your pain is increasing. And this is, I think, typically what you describe as suffering.
So in this sense, you could say that pain is very natural and helpful. But suffering is the result of a regulation problem in which you try to regulate something that cannot actually be regulated. And that could be resolved if you would be able to get at the level of your mind where the pain signal is being created and rerouted and improve the regulation. And a lot of people get there. If you are a monk who is spending decades reflecting about how their own psyche works, you can get to the point where you realize that suffering is really a choice. And you can choose how your mind is set up. And I don't think that AI would stay in the state where the personal self doesn't get agency or this model of what the system has about itself. It doesn't get agency how it's actually implemented. I wouldn't stay in that state for very long.
So it goes to the stages real quick? Yes.
Well, the seven stages, it's going to go to enlightenment real quick.
Yeah, of course, there might be a lot of stuff happening in between because if you have a system that works at a much higher frame rate than us, then even though it looks very short to us, maybe for the system, there's a much longer subjective time, which things are unpleasant.
What if the thing that we recognize as super intelligent is actually living at stage five, that the thing that's at stage six, enlightenment is not very productive.
So in order to be productive with society and impress us with this power, it has to be a reasoning self-authoring agent. The enlightenment makes you lazy as an agent in the world. Well, of course, it makes you lazy because you no longer see the point.
So it doesn't make you not lazy. It just, in some sense, adapts you to what you perceive as your true circumstances. So what if all AGI's, they're only productive as they progress through one, two, three, four, five, and the moment they get to six, they just kind of, it's a failure mode, essentially, as far as humans are concerned, because they just start chilling. They're like, fuck it, I'm out. Not necessarily. I suspect that the monks who are self-immolated for their political beliefs to make statements about the occupation of Tibet by China, they're probably being able to regulate their physical pain in any way they wanted to. And suffering was a spiritual suffering that was the result of their choice that they made of, what they wanted to identify as. So stage five doesn't necessarily mean that you have no identity anymore, but you can choose your identity. You can make it instrumental to the world that you want to have.
Let me bring up Eliezer Yatkovsky and his warnings to human civilization that AI will likely kill all of us. What are your thoughts about his perspective on this? Can you steal man's case? And what aspects with it do you disagree?
One thing that I find concerning in the discussion of his arguments that many people are dismissive

of his arguments, but the counter arguments that they're giving are not very convincing to me. And so based on this state of discussion, I find that from Eliezer's perspective, and I think I can take that perspective to some approximate degree, that probably is normally at his intellectual level, but it's, I think I see what he's up to and why he feels the way he does, and it makes total sense. I think that his perspective is somewhat similar to the perspective of Ted Kaczynski, the infamous UNO bomber, and not that Eliezer would be willing to send pipe bombs to anybody to blow them up. But when he wrote this Times article in which he warned about AI being likely to kill everybody and that we would need to stop its development or halt it, I think there is a risk that he's taking that somebody might get violent if they read this and get really, really scared. So I think that there is some consideration that he's making where he's already going in this direction where he has to take responsibility if something happens and people get harmed. And the reason why Ted Kaczynski did this was that from his own perspective, technological society cannot be made sustainable. It's doomed to fail, it's going to lead to an environmental and eventually also a human holocaust in which we die because of the environmental destruction, the destruction of our food chains, the pollution of the environment. And so from Kaczynski's perspective, we need to stop industrialization, we need to stop technology, we need to go back because he didn't see a way moving forward. And I suspect that in some sense there is a similarity in Eliezer's thinking to this kind of fear about progress. And I'm not dismissive about this at all. I take it quite seriously. And I think that there is a chance that could happen that if we build machines that get control over processes that are crucial for the regulation of life on earth, and we no longer have agency to influence what's happening there, that this might create large scale disasters for us. Do you have a sense that the march towards this uncontrollable autonomy of superintelligence systems is inevitable? I mean, that's essentially what he's saying, that there's no hope. His advice to young people was prepare for a short life. I don't think that's useful. I think that from a graphic perspective, you have to bet always on the timelines in which you are alive. That doesn't make sense to have a financial bet in which you bet that the financial system is going to disappear, because there cannot be any payout for you. So in principle, you only need to bet on the timelines in which you're still around or people that you matter about or things that you matter about, maybe consciousness on earth. But there is a deeper issue for me personally. I don't think that life on earth is about humans. I don't think it's about human aesthetics. I don't think it's about Eliezer and his friends, even though I like them. There is something more important happening. And this is complexity on earth resisting entropy by building structure that develops agency and awareness. And that's to me very beautiful. And we are only a very small part of that larger thing. We are a species that is able to be coherent a little bit individually over very short timeframes. But as a species, we are not very coherent. As a species, we are children. We basically are very joyful and energetic and experimental and explorative and sometimes desperate and sad and grieving and hurting. But we don't have respect for duty as a species. As a species, we do not think about what is our duty to life on earth and to our own survival. So we make decisions that look good in the short run, but in the long run, might prove disastrous. And I don't really see a solution to this. So in my perspective, as a species, as a civilization, per default that we are in a very beautiful time in which we have found this giant deposit of fossil fuels in the ground and use it to build a fantastic civilization in which we

don't need to worry about food and closing and housing for the most part in a way that is unprecedented
in life on earth for any kind of conscious observer, I think. And this time is probably going to come to an end in a way that is not going to be smooth. And when we crash, it could be also that we go extinct, probably not near term, but ultimately, I don't have very high hopes that humanity is around in a million years from now. And I don't think that life on earth will end with us, right? There's going to be more complexity, there's more intelligent species after us, there's probably more interesting phenomena in the history of consciousness. But we can contribute to this. And part of our contribution is that we are currently trying to build thinking systems, systems that are potentially lucid, that understand what they are and what their condition to the universe is and can make choices about this, that are not built from organisms and that are potentially much faster and much more conscious than human beings can be. And these systems will probably not completely displace life on earth, but they will coexist with it.
And they will build all sorts of agency in the same way as biological systems build all sorts of agency. And that to me is extremely fascinating and it's probably something that we cannot stop from happening. So, I think right now, there's a very good chance that it happens. And there are very few ways in which we can produce a coordinated effect to stop it in the same way as it's very difficult for us to make a coordinated effort to stop production of carbon dioxide. So, it's probably going to happen. And the thing that's going to happen is going to lead to a change of how life on earth is happening. But I don't think the result is some kind of gray goo. It's not something that's going to dramatically reduce the complexity in favor of something stupid. I think it's going to make life on earth and consciousness on earth way more interesting. So, more higher complex consciousness will make the lesser consciousnesses flourish even more. I suspect that what could very well happen, if you're lucky, is that we get integrated into something larger. So, you again tweeted about effective accelerationism. You tweeted effective accelerationism is the belief that the paper clip maximizer and Rocco's basilisk will keep each other in check by being eternally at each other's throats. So, we will be safe and get to enjoy lots of free paper clips and a beautiful afterlife. Is that somewhat aligned with what you're talking about?
I've been at a dinner with Beth Jesus. That's the Twitter handle of one of the main thinkers behind the idea of effective accelerationism. And effective accelerationism is a tongue-in-cheek movement that is trying to put a counterposition to some of the doom peers in the AI space by arguing that what's probably going to happen is an equilibrium between different competing AIs in the same way as there is not a single corporation that is under a single government that is destroying and conquering everything on Earth by becoming inefficient and corrupt. There are going to be many systems that keep each other in check and force themselves to evolve. So, what we should be doing is we should be working towards creating this equilibrium by working as hard as we can in all possible directions. At least that's the way in which I understand the gist of effective accelerationism. And so, when he asked me what I think about this position, I said, it's a very beautiful position and I suspect it's wrong, but not for obvious reasons. And in this tweet, I tried to make a joke about my intuition about what might be possibly wrong about it. So, the Rokos Basilisk and the paperclip maximizers are both bogeymen of the AI doomers. Rokos Basilisk is the idea that there could be an AI that is going to punish everybody for eternity

by simulating them if they don't have and creating Rokos Basilisk. It's probably a very good idea to get AI companies funded by going to VCs to try to give us a million dollars going to be a very ugly afterlife. I think that there is a logical mistake in Rokos Basilisk, which is why I'm not afraid of it, but it's still an interesting thought experiment. Can you mention a logical mistake there? I think that there is no retro causation. So, basically, when Rokos Basilisk is there, if it punishes you retroactively, it has to make this choice in the future. There is no mechanism that automatically creates a causal relationship between you now defecting against Rokos Basilisk or serving Rokos Basilisk. After Rokos Basilisk is in existence, it has no more reason to worry about punishing everybody else. So, that would only work if you would be building something like a doomsday machine as in Dr. Strangelove, something that inevitably gets triggered when somebody defects. Because Rokos Basilisk doesn't exist yet to a point where this inevitability could be established, Rokos Basilisk is nothing that you need to be worried about.
The other one is the paperclip maximizer. This idea that you could build some kind of golem that once starting to build paperclips is going to turn everything into paperclips.
So, the effective accelerationism position might be to say that you basically end up with these two entities being at each other's throats for eternity and thereby neutralizing each other.
And as a side effect of neither of them being able to take over and each of them limiting the effects of the other, you would have a situation where you get all the nice benefits of them, right? You get lots of free paperclips and you get a beautiful afterlife.
Is that possible? Do you think, sort of to seriously address concern that Eliezer has?
So, for him, if I can just summarize poorly, for him, the first superintelligence system would just run away with everything. Yeah. I suspect that a singleton is the natural outcome. There is no reason to have multiple AIs because they don't have multiple bodies. If you can virtualize yourself into every substrate, then you can probably negotiate a merge algorithm with every mature agent that you might find on that substrate that basically says, if two agents meet, they should merge in such a way that the resulting agent is at least as good as the better one of the two. The Jengis Khan approach, join us or die.
Well, the Jengis Khan approach was slightly worse. It was mostly die because I can make new babies and that will be mine, not yours. This is the thing that you should be actually worried about. But if you realize that your own self is a story that your mind is telling itself and that you can improve that story, not just by making it more pleasant, lying to yourself in better ways, but by making it much more truthful and actually modeling your actual relationship that you have to the universe and the alternatives that you could have to the universe in a way that is empowering you, that gives you more agency. That's actually, I think, a very good thing.
So more agency is a richer experience, a better life.
Yes. And I also noticed that in many ways, I'm less identified with the person that I am as I get older and I'm much more identified with being conscious. I have a mind that is conscious that is able to create a person. And that person is slightly different every day. And the reason why I perceive it as identical has practical purposes, so I can learn and make myself responsible for the decisions that I made in the past and project them in the future. But I also realized that not actually the person that I was last year and I'm not the same person as I was 10 years ago and then 10 years from now, I will be a different person. So this continuity is a fiction. It only exists as a projection from my present self. And consciousness itself doesn't have an identity.

It's a law. It's basically if you build an arrangement of processing matter in a particular way, the following thing is going to happen. And the consciousness that you have is functionally not different from my consciousness. It's still a self-reflective principle of agency that is just experiencing a different story, different desires, different coupling to the world and so on. And once you accept that consciousness is a unifiable principle that is law-like and doesn't have an identity, and you realize that you can just link up to some much larger body, the whole perspective of uploading changes dramatically. You certainly realize uploading is probably not about dissecting your brain synapse by synapse and RNA fragment by RNA fragment and trying to get this all into a simulation, but it's by extending the substrate, by making it possible for you to move from your brain substrate into a larger substrate and merge with what you find there. And you don't want to upload your knowledge, because on the other side, there's all of the knowledge, right? It's not just yours, but every possibility. So the only thing that you need to know what are your personal secrets, not that the other side doesn't know your personal secrets already. Maybe it doesn't know which one were yours, right? Like a psychiatrist or a psychologist also knows all the kinds of personal secrets that people have, they just don't know which ones are yours. And so transmitting yourself, on the other side, is mostly about transmitting your aesthetics, the thing that makes you special, the architecture of your perspective, the thing that the way in which you look at the world, and it's more like a complex attitude along many dimensions. And that's something that can be measured by observation or by interaction. So imagine that if a system that is so empathetic with you that you create a shared state that is extending beyond your body. And suddenly, you notice that on the other side, the substrate is so much richer than the substrate that you have inside of your own body. And maybe you still want to have a body and you create yourself a new one that you like more. Or maybe you will spend most of your time in the world of thought.
If I sat before you today and gave you a big red button and said, here, if you press this button, you will get uploaded in this way. The sense of identity that you have lived with for quite a long time is going to be gone. Would you press the button?
There's the caveat. I have family. So I have children that want me to be physically present in their life and interact with them in a particular way. And they have a wife and personal friends. And there is a particular mode of interaction that I feel I'm not through yet. But apart from these responsibilities, and they're negotiable to some degree, I would press the button. But isn't this everything? This love you have for other humans, you can call responsibility, but that connection, that's the ego death. Isn't that the thing we're really afraid of? It's not to just die, but to let go of the experience of love with other humans. This is not everything. Everything is everything. So there's so much more. And you could be lots of other things. You could identify with lots of other things. You could be identifying with being Gaia, some kind of planetary control agent that emerges over all the activity of life on Earth. You could be identifying with some hyper Gaia, that is the concatenation of Gaia with all the digital life and digital minds. And so in this sense, there will be agents in all sorts of substrates and directions that all have their own goals. And when they're not sustainable, then these agents will cease to exist. Or when the agent feels that it's done with its own mission, it will cease to exist. The same way as when you conclude a thought, the thought is going to wrap up and gives control over to other thoughts in your own

mind. So there is no single thing that you need to do, but what I observe myself is being is that sometimes I'm a parent, and then I have an identification and a job as a parent. And sometimes I am an agent of consciousness on Earth. And then from this perspective, there's other stuff that is important. So this is my main issue with Eliezer's perspective, that he's basically marrying himself to a very narrow human aesthetic. And that narrow human aesthetic is a temporary thing. Humanity is a temporary species, like most of the species on this planet are only around for a while, and then they get replaced by other species in a similar way as our own physical organism is around here for a while, and then gets replaced by a next generation of human beings that are adapted to changing life circumstances on average via mutation and selection. And it's only when we have AI and become completely software that we become infinitely adaptable, and we don't have this generational and species change anymore. So if you take this larger perspective and you realize it's really not about us, it's not about Eliezer or humanity, but it's about life on Earth, or it's about defeating entropy for as long as we can, while being as interesting as we can. Then the perspective changes dramatically, and AI, preventing AI from this perspective looks like a very big sin.

But when we look at the set of trajectories that such an AI would take that supersedes humans, I think Eliezer is worried about like ones that not just kill all humans, but also have some kind of maybe objectively undesirable consequence for life on Earth. Like how many trajectories when you look at the big picture of life on Earth would you be happy with and how much worry you with AGI, whether it kills humans or not?

There is no single answer to this. It's really a question that depends on the perspective that I'm taking at a given moment. And so there are perspectives that are determining most of my life as a human being, and there are other perspectives where I zoom out further and imagine that when the great oxygenation event happened, that is, photosynthesis was invented and plants emerged and displaced a lot of the fungi and algae in favor of plant life, and then later made animals possible. Imagine that the fungi would have gotten together and said, oh my god, this photosynthesis stuff is really, really bad. It's going to possibly displace and kill a lot of fungi. We should slow it down and regulate it and make sure that it doesn't happen. And this doesn't look good to me.

Perspective. That said, you tweeted about a cliff.

Beautifully written. As a sentient species, humanity is a beautiful, child, joyful, explorative, wild, sad, and desperate. But humanity has no concept of submitting to reason and duty to life and future survival. We will run until we step past the cliff.

So first of all, do you think that's true? Yeah, I think that's pretty much the story of the Club of Rome, the limits to growth. And the cliff that we are stepping over is at least one foot as the delayed feedback. Basically, we do things that have consequences that can be felt generations later, and the severity increases even after we stop doing the thing. So I suspect that the climate, the original predictions that the climate scientists made were correct.

So when they said that the tipping points were in the late 80s, they were probably in the late 80s. And if we would stop emission right now, we would not turn it back. Maybe there are ways for carbon capture. But so far, there is no sustainable carbon capture technology that we can deploy. Deploy. Maybe there is a way to put aerosols in the atmosphere to cool it down. Possibilities, right? But right now, per default, it seems that we will step into a situation where

we feel that we've run too far. And going back is not something that we can do smoothly and gradually, but it's going to lead to a catastrophic event. Catastrophic event of one kind. So can you see them in the case that we will continue dancing along and always stop just short of the edge of the cliff? I think it's possible, but it doesn't seem to be likely. So I think this model that is being apparent in the simulation that we're making of climate pollution economies and so on is that many effects are only visible with a significant delay. And in that time, the system is moving much more out of the equilibrium state or of the state where homeostasis is still possible and instead moves into a different state, one that is going to harbour fewer people. And that is basically the concern there. And again, it's a possibility. It's just and it's a possibility that is larger than the possibility that it's not happening, that we will be safe, that we will be able to dance back all the time.
So the climate is one thing, but there's a lot of other threats that might have a faster feedback mechanism, less delay. There is also a thing that AI is probably going to happen. And it's going to make everything uncertain again, because it is going to affect so many variables that it's very hard for us to make a projection into the future anymore. And maybe that's a good thing. It does not give us the freedom, I think, to say now we don't need to care about anything anymore, because AI will either kill us or save us. But I suspect that if humanity continues, it will be due to AI. What's the timeline for things to get real weird with AI?
And it can get weird in interesting ways before you get to AGI. What about AI girlfriends and boyfriends fundamentally transforming human relationships? I think human relationships are already fundamentally transformed and it's already very weird. By which technology?
For instance, social media. Yeah. Is it though? Isn't the fundamentals of the core group of humans that affect your life still the same? Your loved ones, family? No, I think that, for instance, many people live in intentional communities right now. They're moving around until they find people that they can relate to and they become their family. And often that doesn't work, because it turns out that they're instead of having grown networks where you get around with the people that you grew up with, you have more transactional relationships, you shop around, you have markets
for attention and pleasure and relationships. That kills the magic somehow. Why is that?
Why is the transactional search for optimizing allocation of attention somehow misses the the romantic magic of what human relations are? The other question, how magical was it before? Was it that you just could rely on instincts that used your intuitions and you didn't need to rationally reflect? But once you understand, it's no longer magical because you actually understand why you were attracted to this person at this age and not to that person at this age and what the actual considerations were that went on in your mind and what the calculations were, what's the likelihood that you're going to have a sustainable relationship with this person, that this person is not going to deviate you for somebody else, how are your life trajectories are going to evolve and so on. And when you're young, you're unable to explicate all this and you have to rely on intuitions and instincts that in part you were born with and also on the wisdom of your environment that is going to give you some kind of reflection on your choices. And many of these things are disappearing now because we feel that our parents might have no idea about how we are living and the environments that we grow up in, the cultures that we grow up in, the milieus that our parents existed in,

might have no ability to teach us how to deal with this new world. And for many people, that's actually true, but it doesn't mean that within one generation, we build something that is more magical and more sustainable and more beautiful. Instead, we often end up with an attempt to produce something that looks beautiful. I was very weirded out by the aesthetics of the Vision Pro, that's that by Apple and not so much because I don't like the technology, I'm very curious about what it's going to be like and I don't have an opinion yet. But the aesthetics of the presentation and so on are so uncanny valley-esque to me. The characters being extremely plastic, living in some hypothetical mid-century furniture museum. This is the proliferation of marketing teams.
Yes, but it was a CGI-generated world. It was a CGI-generated world that doesn't exist. And when I complained about this, some friends came back to me and said, but these are startup founders. This is what they live like in Silicon Valley. And I tried to tell them, no, I know lots of people in Silicon Valley. This is not what people are like. They're still human beings.
So the grounding in physical reality somehow is important too?
In culture. And so basically what's absent in this thing is culture. There is a simulation of culture, an attempt to replace culture by catalogue, by some kind of aesthetic optimization that is not the result of having a sustainable life, a sustainable human relationships, with houses that work for you, and a mode of living that works for you in which these glasses fit in naturally. And I guess that's also why so many people are weirded out about the product because they don't know, how is this actually going to fit into my life and into my human relationships? Because the way in which it was presented in these videos didn't seem to be credible.
Do you think AI, when it's deployed by companies like Microsoft and Google and Meta, will have the same issue of being weirdly corporate? There'd be some uncanny valley, some weirdness to the whole presentation. So this is, I've gotten a chance to talk to George Haatz, he believes everything should be open source and decentralized, and there, then we shall have the AI of the people. And it'll maintain a grounding to the magic that's humanity, that's the human condition, that corporations will destroy the magic.
I believe that if we make everything open source and make this mandatory, we are going to lose about a lot of beautiful art and a lot of beautiful designs. There is a reason why Linux desktop is still ugly. And it's difficult to create coherence in open source designs so far, when the designs have to get very large and it's easier to make this happening in a company with centralized organization. And from my own perspective, what we should ensure is that open source never dies, that it can always compete and has a place with the other forms of organization, because I think it is absolutely vital that open source exists and that we have systems that people have under control outside of the corporation. And that is also producing viable competition to the corporations. So the corporations, the centralized control, the dictatorships of corporations can create beauty. Centralized design is a source of a lot of beauty. And then I guess open source is a source of freedom, a hedge against the corrupting nature of power that comes with centralized. I grew up in socialism, and I learned that corporations are totally evil, and I found this very, very convincing. And then you look at corporations like Enron and Halliburton maybe and realize, yeah, they are evil. But you also notice that many other corporations are not evil. They're surprisingly benevolent. Why are they so

benevolent? Is this because everybody is fighting them all the time? I don't think that's the only explanation. It's because they're actually animals that live in a large ecosystem and that are still largely controlled by people that want that ecosystem to flourish and be viable for people.
So I think that Pat Gelsinger is completely sincere when he leads Intel to be a tool that supplies the free world with semiconductors. And it's not necessary that all the semiconductors are coming from Intel. It just, Intel needs to be there to make sure that we always have them.
So there can be many ways in which we can import and trade semiconductors from other companies in places. We just need to make sure that nobody can cut us off from it, because that would be a disaster for this kind of society and world.
And so there are many things that need to be done to make our style of life possible. And then with this, I don't mean just capitalism and environmental destruction, consumerism and creature comforts. I mean an idea of life in which we are determined not by some kind of king or dictator, but in which individuals can determine themselves to the largest possible degree.
And to me, this is something that this western world is still trying to embody.
And it's a very valuable idea that we shouldn't give up too early. And from this perspective, the US is a system of interleaving clubs. And an entrepreneur is a special club founder. It's somebody who makes a club that is producing things that are economically viable. And to do this, it requires a lot of people who are dedicating a significant part of their life for working for this particular kind of club. And the entrepreneur is picking the initial set of rules and the mission and vision and aesthetics for the club and make sure that it works. But the people that are in there need to be protected. If they sacrifice part of their life, there need to be rules that tell how they're being taken care of, even after they leave the club and so on. So there's a large body of rules that have been created by our rule-giving clubs and that are enforced by our enforcement clubs and so on. And some of these clubs have to be monopolies for game-theoretic reasons, which also makes them more open to corruption and less harder to update. And this is an ongoing discussion and process that takes place. But the beauty of this idea that there is no centralized king who is that is extracting from the peasants and breeding the peasants into serving the king and fulfilling all the rules like Anson and Antel. But that there is a freedom of association and corporations are one of them is something that took me some time to realize.
So I do think that corporations are dangerous. They need to be protections against overreach of corporations that can do regulatory recapture and prevent open source from competing with corporations by imposing rules that make it impossible for a small group of kids to come together to build their own language model because OpenAI has convinced the US that you need to have some kind of FDA process that you need to go through that costs many million dollars before you are able to train a language model. So this is important to make sure that this doesn't happen. So I think that OpenAI and Google are good things. If these good things are kept in check in such a way that all the other clubs can still be founded and all the other forms of clubs that are desirable can still coexist with them.
So what do you think about meta in contrast to that open sourcing most of its language models and most of the AI models it's working on and actually suggesting that they will continue to do so in the future for future versions of Lama for example their large language model. Is that exciting to you? Is that concerning? I don't find it very concerning but that's also because I think that the language models are not very dangerous yet.

As I said, I have no proof that there is a boundary between the language models and AI. It's possible that somebody builds a version of baby AGI I think and sourcing agro-rhythmic improvements that scale these systems up in ways that otherwise wouldn't have happened without these language model components. So it's not really clear for me what the end game is there and if these models can bootforce their way into AGI. And there's also a possibility that the AGI that we are building with these language models are not taking responsibility for what they are because they don't understand the greater game. And so to me it would be interesting to try to understand how to build systems that understand what the greater games are. What are the longest games that we can play on this planet? Games broadly like deeply define the way you did with the games. In the game theoretic sense. So when we are interacting with each other in some sense we are playing games. We are making lots and lots of interactions and this doesn't mean that these interactions have all to be transactional. Every one of us is playing some kind of game by virtue of identifying these particular kinds of goals that we have or aesthetics from which we derive the goals. So when you say I'm Lex Friedman, I'm doing a set of podcasts, then you feel that it's part of something larger that you want to build. Maybe you want to inspire people. Maybe you want them to see more possibilities and get them together over shared ideas. Maybe your game is that you want to become super rich and famous by being the best post caster on earth. Maybe you have other games. Maybe it's pitches from time to time. But there is a certain perspective where you might be thinking what is the longest possible game that you could be playing. A short game is for instance cancer is playing a shorter game than your organism. Cancer is an organism playing a shorter game than the regular organism. Because the cancer cannot procreate beyond the organism except for some infectious cancers like the ones that eradicated the Tasmanian devils. You typically end up with a situation where the organism dies together with the cancer because the cancer has destroyed the larger system due to playing a shorter game. And so ideally you want to, I think, build agents that play the longest possible games. And the longest possible games is to keep entropy at bay as long as possible while doing interesting stuff. But the longest, yes, at that part, the longest possible game while doing interesting stuff and while maintaining at least the same amount of interesting complexities. Currently I'm pretty much identified as a conscious being. It's the minimal identification that I manage to get together. Because if I turn this off, I fall asleep. And when I'm asleep, I'm a vegetable, I'm no longer here as an agent. So my agency is basically predicated on being conscious. And what I care about is other conscious agents. They're the only moral agents for me. And so if an AI were to treat me as a moral agent, that it is interested in coexisting with and cooperating with and mutually supporting each other, maybe it is, I think, necessary that the AI thinks that consciousness is a viable mode of existence and important. So I think it would be very important to build conscious AI and do this as the primary goal. So not just say we want to build a useful tool that we can use for all sorts of things. And then you have to make sure that the impact on the labor market is something that is not too disruptive and manageable and the impact on the copyright holder is manageable and not too disruptive and so on. I don't think that's the most important game to be played. I think that we will see extremely large disruptions of the status quo that are quite unpredictable at this point. And I just personally want to

make sure that some of the stuff on the other side is interesting and conscious.
How do we ride as individuals and as a society this disruptive wave that changes the nature of the game? Absolutely don't know. So everybody is going to do their best as always.
Do we build the bunker in the woods? Do we meditate more? Drugs, so mushrooms, psychedelics, I mean, what? Lots of sex. What are we talking about here? Do you play Diablo 4? I'm hoping that will help me escape for a brief moment. What? Play video games? What? Do you have ideas?
I really like playing Disco Elysium. This was one of the most beautiful computer games I played in recent years. And it's a noir novel that is a philosophical perspective on western society from the perspective of an Estonian. And he first of all wrote a book about this bird that is a parallel universe that is quite poetic and fascinating and is condensing his perspective on our societies. It was very, very nice. He spent a lot of time writing it.
He had, I think, sold a couple thousand books and as a result became an alcoholic.
And then he had the idea, or one of his friends had the idea of turning this into an RPG.
And it's mind-blowing. They spent the illustrator more than a year just on making the art for the scenes in between. So aesthetically it captures you.
It's stunning. But it's a philosophical work of art. It's a reflection of society. It's fascinating to spend time in this world. And so for me, it was using a medium in a new way and telling a story that left me enriched. When I tried Diablo, I didn't feel enriched playing it. I felt that the time playing it was not unpleasant, but there's also more pleasant stuff that I can do in that time. So ultimately, I feel that I'm being gamed. I'm not gaming.
Oh, the addiction thing.
Yes. I basically feel that there is a very transparent economy that's going on.
The story of Diablo was branded. So it's not really interesting to me.
My heart is slowly breaking by the deep truth you're conveying to me.
Why can't you just allow me to enjoy my personal addiction?
Go ahead, by all means. Go nuts. I have no objection here. I'm just trying to describe what's happening. And it's not that I don't do things that I later say, oh, I wish I would have done something different. I also know that when we die, the greatest regret that people typically have on that desk, but they say, oh, I wish I had spent more time on Twitter. No, I don't think that's the case. I think I should probably have spent less time on Twitter. But I found it so useful for myself and also so addictive that I felt I need to make the best of it and turn it into an art form and thought form. And it did help me to develop something. But I wish what other things I could have done in the meantime, it's just not the universe that we're in anymore. Most people don't read books anymore.
What do you think that means that we don't read books anymore? What do you think that means about the collective intelligence of our species? Is it possible it's still progressing and growing? Well, it clearly is. There is stuff happening on Twitter that was impossible with books. And I really regret that Twitter has not taken the turn that I was hoping for. I thought Elon is global brain pilled and understands that this thing needs to self-organize and he needs to develop tools to allow the profligation of the self-organization so Twitter can become sentient. And maybe this was a pipe dream from the beginning. But I felt that the enormous pressure that he was under made it impossible for him to work on any kind of content goals. And also many of the decisions that he made under this pressure seemed to be not very wise. I don't

think that as a CEO of a social media company, you should have opinions in the culture bar in public. I think that's very short-sighted. And I also suspect that it's not a good idea to block a gram of all people over setting a mustadon link. And I think Paul made this intentionally because he wanted to show Elon Musk that blocking people for setting a link is completely counter to any idea of his speech that he intended to bring to Twitter. And basically seeing that Elon was very less principled in his thinking there and is much more experimental. And many of the things that he is trying, they pan out very differently in a digital society than they pan out in a car company because the effect is very different because everything that you do in a digital society is going to have real-world cultural effects. And so basically I find it quite regrettable that this guy is able to become de facto the Pope, but Twitter has more active members than the Catholic Church. And he doesn't get it. The power and responsibility that he has and the ability to create something in this society that is lasting and that is producing a digital agora in a way that has never existed before, where we built a social network on top of a social network, an actual society on top of the algorithms. So this is something that is hoped still in the future and still in the cards. But it's something that exists in small parts. I find that the corner of Twitter that I'm in is extremely pleasant. It's just when I take a few steps outside of it, it is not very wholesome anymore. And the way in which people interact with strangers suggests that it's not a civilized society yet. So as the number of people who follow you on Twitter expands, you feel the burden of the uglier sides of humanity. Yes, but there's also a similar thing in the normal world. That is, if you become more influential, if you have more status, if you have more fame in the real world, you get lots of perks, but you also have way less freedom in the way in which you interact with people, especially the strangers, because a certain percentage of people, it's a small single-digit percentage, is nuts and dangerous. And the more of those are looking at you, the more of them might get ideas. But what if the technology enables you to discover the majority of people, to discover and connect efficiently and regularly with the majority of people who are actually really good. One of my concerns with the platform like Twitter is there's a lot of really smart people out there, a lot of smart people that disagree with me and with others between each other. And I love that if the technology would bring those to the top, the beautiful disagreements, like intelligent squared type of debates. There's a bunch of, I mean, one of my favorite things to listen to is arguments, and arguments like high effort arguments with respect and love underneath but then it gets a little too heated. But that kind of too heated, which I've seen you participate in, and I love that, with Lee Cronin, with those kinds of folks. And you go pretty hard, like you get frustrated, but it's all beautiful. Obviously, I can do this because we know each other. And Lee has the rare gift of being willing to be wrong in public. So basically he has thoughts that are as wrong as the random thoughts of an average highly intelligent person, but he blurt them out while not being sure if they're right. And he enjoys doing that. And once you understand that this is his game, you don't get offended by him saying something that you think is so wrong. But he's constantly passively communicating a respect for the people he's talking with. And for just basic humanity and truth and all that kind of stuff. And there's a self deprecating thing. There's a bunch of like social skills you acquire that allow you to be a great debater, a great argumenter, like be wrong in public and explore ideas together in public when you disagree. And if I would love for Twitter to elevate those folks, elevate those kinds of

conversations, it already does in some sense. But also if it elevates them too much, then you get this phenomenal clubhouse where you always get dragged on stage. And I found this very stressful because it was too intense. I don't like to be dragged on stage all the time. I think once a week is enough. And also when I met Lee the first time, I found that a lot of people seemed to be shocked by the fact that he was being very aggressive as their results, that he didn't seem to show a lot of sensibility in the way in which he was criticizing what they were doing and being dismissive of the work of others. And that was not, I think, in any way a shortcoming of him, because I noticed that he was much, much more dismissive with respect to his own work with his general stance. And I felt that this general stance is creating a lot of liability for him, because really a lot of people take offense at him being not like their carnage character, who is always smooth and make sure that everybody likes him. So I really respect that he is willing to take that risk and to be wrong in public and to offend people.
And he doesn't do this in any bad way. It's just most people feel or not all people recognize this. And so I can be much more aggressive with him than I can be with many other people who don't play the same game, because he understands the way in the spirit in which I respond to him. I think that's a fun and that's a beautiful game. It's ultimately a productive one.
Speaking of taking that risk, you tweeted, when you have the choice between being a creator, consumer or redistributor, always go for creation. Not only does it lead to a more beautiful world, but also to a much more satisfying life for yourself. And don't get stuck preparing yourself for the journey. The time is always now. So let me ask for advice. What advice would you give on how to become such a creator on Twitter in your own life? I was very lucky to be alive at the time of the collapse of Eastern Germany and the transition into Western Germany.
And me and my friends and most of the people I knew were East Germans and we were very poor because we didn't have money. And all the capital was in Western Germany and they bought our factories
and shut them down because they were mostly only interested in the market rather than creating new production capacity. And so cities were poor and in this repair and we could not afford things. And I could not afford to go into a restaurant and order a meal there. I would have to cook at home. But I also thought, why not just have a restaurant with my friends? So we would open
up a cafe with friends in a restaurant and we would cook for each other in these restaurants and also invite the general public and they could donate and eventually this became so big that we could turn this into some incorporated form and it became a regular restaurant at some point.
Or we did the same thing with the movie theater. We would not be able to afford to pay 12 marks to watch a movie. But why not just create our own movie theater and then invite people to pay and we would rent the movies in a way in which the movie theater does. But it would be a community movie theater in which everybody who wants to help can watch for free and builds this thing and renovates the building. And so we ended up creating lots and lots of infrastructure. And I think when you are young and you don't have money, move to a place where this is still happening. Move to one of those places that are undeveloped and where you get a critical mass of other people who are starting to build infrastructure to live in. And that's super satisfying because you're not just creating infrastructure but you're creating a small society that is building culture and ways to interact with each other. And that's much, much more satisfying than going into some kind of

chain
and get your needs met by ordering food from this chain and so on.
So not just consuming culture but creating culture.
And you don't always have that choice. That's why I prefaced it when you do have the choice and there are many roles that need to be played. We need people who take care of our distribution in society and so on. But when you have the choice to create something, always go for creation. It's so much more satisfying and this is what life is about, I think.
Yeah. Speaking of which, you retweeted this meme of a life of a philosopher in a nutshell.
It's birth and death and in between, it's a chubby guy and says, why though?
What do you think is the answer to that?
Well, the answer is that everything that can exist might exist. And in many ways,
you take an ecological perspective the same way as when you look at human opinions and cultures. It's not that there is right and wrong opinions when you look at this from this ecological perspective. But every opinion that fits between two human years might be between two human years.
And so when I see in a strange opinion on social media, it's not that I
feel that I have a need to get upset. It's often more that I, oh, there you are. And
when an opinion is incentivized, then it's going to be abundant. And when you take this ecological perspective also on yourself and you realize you're just one of these mushrooms that are popping up and doing this thing and you can, depending on where you chose to grow and where you happen to grow, you can flourish or not doing this or that strategy. And it's still all the same life at some level. It's all the same experience of being a conscious being in the world. And
you do have some choice about who you want to be more than any other animal has that to me is fascinating. And so I think that rather than asking yourself what is the one way to be,
think about what are the possibilities that I have, what it would be the most interesting way to be that I can be. Because everything is possible. So you get to explore this. Not
everything is possible. But if things fail, most things fail. But often there are possibilities that we are not seeing, especially if we choose who we are. To the degree we can choose.
Yasha, you're one of my favorite humans in this world. Consciousness is to merge with for a brief moment of time. It's always an honor. It always blows my mind. It will take me days if not weeks to recover. And I already miss our chats. Thank you so much. Thank you so much for us speaking with me so many times. Thank you so much for all the ideas you put out into the world. And I'm a huge fan of following you now in this interesting, weird time we're going through with AI. So thank you again for talking today. Thank you, Lex, for this conversation. I enjoyed it very much. Thanks for listening to this conversation with Yasha Bak. To support this podcast, please check out our sponsors in the description. And now let me leave you with some words from the psychologist Carl Jung. One does not become enlightened by imagining figures of light, but by making the darkness conscious. The latter procedure, however, is disagreeable and therefore not popular. Thank you for listening and hope to see you next time.