POD OF JAKE
#157 - VITALIK BUTERIN
AI-GENERATED TRANSCRIPT

Jake 00:16
All right, thank you, Vitalik, for joining me on the podcast today. I really appreciate it. You were kind enough to be one of the very first guests on the podcast a few years ago. And it's great to be able to talk to you again, I reached out, you know, a little while ago, wanted to talk about Zulu. And, you know, at the time, the open AI, Sam Altman sort of dispute was very hot and unresolved at that point. And I want to get your perspective on that. Because I know historically, you've been very interested in somewhat involved with the Effective Altruism movement. And now, you know, we've got yak and those two sort of butted heads with this dispute in a way that I found to be sort of frustratingly toxic, and tribal. And I thought you might have sort of a nuanced view, and then, you know, so that's what I wanted to talk about. And then to my delight, you actually put out a piece called my technical optimism, which, you know, very clearly shared your perspective. And I think that was just great for, for the issue overall, and for people paying attention, because it was much more nuanced than, you know, a little tweet of, let's go IAQ, or, you know, stop ball AI or anything like that. So what I'd like to do today, as we discussed before, is to just sort of like walk through your piece a little bit, but hopefully, you know, talk about a lot of questions that you didn't explicitly address. So people can go and read it, you know, beforehand, or, or after the fact and hopefully get something additional, you know, hopefully a lot additional from from this conversation as well. So I guess the best place to start would be what, what motivated you to write this piece in the first place, was it some combination of, you know, the increasing yak movement, or, you know, Marc Andreessen piece, which you mentioned, or the dispute, or kind of all the above or something else?

Vitalik Buterin 01:57
Yeah. And so I started thinking about a lot of these ideas, like basically, actually, in the week that Marc Andreessen zoo piece came out, this was back in October. And I remember reading it, and I remember seeing the responses to it. And I remember getting this distinct impression that there's some kind of missing perspective, that hasn't really made it out anywhere in public yet. And so I actually just started thinking about whether or not I should write some kind of piece and just go out and try to write the piece that I

thought should exist myself. And at the same time, he was that as an opportunity to really finish putting a lot of my thoughts on some of these topics together. It was interesting, because a lot of these issues were issues that I had been thinking about in the back of my mind a lot throughout this whole year. And I thought a lot about them, while Zulu was happening in the spring, just because of how effectively that events brought together. Both a lot of crypto people, but also a lot of people interested in AI and bio and some of these other topics. And one of the challenges that was going through my mind is that I noticed that I had these thoughts and opinions on crypto. And then separately, I had these thoughts and opinions on bio. And then separately, I had these thoughts and opinions on Effective Altruism, and morality, and then like ethics, and all of those questions. And then separately, I had my thoughts on AI and the singularity. And they weren't really in, you know, what the cool kids call reflective equilibrium with each other. Right? Like, I didn't feel like I had these separate strands of thoughts that I had where he had consistent or really, uh, yeah. How would I Yeah, put it like that? And like, in the context of each other? Yeah. Yeah. And so I started thinking, thinking in more detail, and just making sure that I was properly updating toward, I'm gonna look what we've seen over the last couple of years in technology, some of the very negative things that we've seen over the last couple of years in politics, some of the trends that we've seen in other spaces, also, the things that I really like and value and respect to and also some of the things that I disrespect in the crypto space, and like basically how all of those things might or fit together into one common picture. And so I started thinking about the piece and generally when I write things, there's like, a period when I write it, which usually Yeah, kind of lasts maybe two or three days. That includes a couple of stuff. batches of a couple of hours each, then there is a period of reaching out for feedback, collecting feedback, thinking about the feedback, and then finally publishing it. But then before that there's definitely a longer period where I spent the time just thinking through those issues. And often, by the time of my keyboard hits to hack MD editor, like three quarters of what I have in mind, this like, kind of figured out in my head already, though, often the process of writing itself does end up having to look revealing important kind of points to having considered or contradictions and so forth. And then it just so happens that the week after this zoo connect was happening right in zoo connect was the Zulu gathering that was meant as a sister events, in some ways to Deaf connect, right. So deaf Connect is the annual Etherium sort of conference sort of unconference, right, so like you might have heard of DEF CON, and the which is the annual Etherium Conference, which is, of course, a parody of DEF CON with an F, which is I'm an org that I'd be even more famous security conference. But then last year, we started doing def connects, which shows like a version of DEF CON about where things are more distributed. And it's more about many independent events, instead of relying on the Ethereum foundation to organize everything. And so last year, there was a def connect and and DEF CON. And then this year, there is a def connect instead of a DEF CON. And next year, there's going to be a DEF CON again, we're still mixing and matching the styles event. But Alongside this, there was a zoo Connect, which was let's use the word community but gathering together for two weeks inside of a city, and basically do having the same kinds of discussions that we had inside of Cisco, and continuing along a lot of those threads. But we're also in a somewhat lighter weight way and not kind of insisting that everyone live in the same place for two whole months. Um, so a lot of the same people were there. And that gave me a lot of opportunities to talk to a lot of interesting people about crypto about AI, also about brain computer interfaces, right. So it was interesting, I had a lot of chats with one Bennett from like file coin and IPFS over the last half year or so. And he mentioned to me how he was really bullish on this concept of like brain computer interfaces, and leading up to essentially merging with AI instead of creating totally separate kind of AI. And I remember I found this direction compelling. And then also in October I had read or I guess, audiobook read the new Elon Musk biography. And I learned there that like Elon was convinced about the exact same idea. And then I learned that a couple of other people were convinced about the exact same idea. And millon, Jevic KOVITCH. Smith, KOVITCH, a researcher who's been really involved in the BCI space, Mike also had a good chance to talk to him. Um, so it was in this kind of environment, where I got to, like, really get to see and understand a lot of these other scientific trends better. And then, of course, all of the ongoing COVID research stuff, which I've been involved in ever since I'm a crypto relief and ball we started a couple of years ago. So started writing and started putting together my own perspective, which is basically, uh, yeah, like, I really believe in the value of technology, as usual. And as I know, lots of your guests do, because I've listened to them. And, you know, okay, yeah, really see it, just like how much technology improves the lives of people. I'm just really massively and how easy that is to under appreciate. And but at the same time, I'm going to be taking a lot of the AI safety arguments seriously. And taking seriously some of the issues that I see in some of the other tech sectors, right. So issues around, you know, computer security issues around some of the things that crypto is trying to solve social media issues, you basically put that together into a common picture. And so this concept of DIAC, roaches sort of a little bit intense intention of ambiguous premium defense, acceleration, but also worth a touch of democracy and, and decentralization is basically where a lot of those ideas came from. And like, that's what the post is about.

Jake 09:49
Great. Well, thank you. For all that context. It's super helpful and sort of wasn't aware of the backstory. So it's really interesting, a couple of things just at a macro level. First of all, I love that you took the time to to write this, obviously, it's like not, it's much easier to like fire off a tweet. But to bring all of these, you know, super complex threads together in your head and kind of write a nuanced piece as you did to, like you said, bring your own, you know, to develop your own perspective, even further, it's like hard to mash these things all together in your head, but through the writing process, and it's sort of like an iterative form of thinking in a way. So you write it, and then you iterate. And then you write it and you iterate, and you actually develop your own thoughts. And then at the end, you have a, you know, the ability to share that in a very coherent way with with the public. So I appreciate you doing that I know, like you took, you're a big fan of long form, I'm a big fan of long form, I think you took the time off of Twitter, where you were just doing long form. So I don't know how to make that like, more popular, because it's just, you have to take the time to, you know, to write it and to read it. But I wish we had more of that sort of, you know, Twitter, if Twitter is sort of the town square, I wish the town square had a little bit more like a system that was built a little bit more for, for nuance. So anyway, when we do have these opportunities,

I really appreciate it. One question that comes to mind early, I think, to your point, you know, we're both I identify as like a techno optimist, you identify as a technical optimist, but it's not so simple as to say like, oh, everything that you know, IAQ supports on like, you know, no question, this is all like, super easy, let's just move forward, like this is all obvious stuff. That's where the tribal kind of stuff starts to come in. And I don't even know, you know, Deac is what you sort of named your perspective, I don't even know if I completely identify with that. And I don't think there's really like, necessarily an issue with that, I think it's definitely the closest to my perspective, and I probably have to take the time to sort of write out my own in order to even know exactly what it is. But nonetheless, I love your view on it. One of the things you sort of point out very early, like that's sort of a, you know, set the stage a little bit is that, you know, you are a very strong believer that technology makes the world a better place. And moreover, that humans are fundamentally good. You talk about, you know, various examples of, you know, one being the doubling of lifespan over the last century, and, and all of these different examples of how tech has made the world better. And you talked about how sort of like, generally, delaying technology, in most cases, can be like very punitive. But are there examples where rushing technology has been negative in the past? Or are you sort of just treating AI as maybe sort of an unprecedented time where accelerating too fast, could be negative to the point where it is, you know, irreversible. And even, you know, you bring up the environment being sort of like the one big exception to the last 100 years where the environment was a negative consequence. But with the environment, no one's claiming, oh, this is going to extinct humanity, or this is completely reversible. And in fact, you express optimism that we can solve that issue just as we solve pollution and other environmental related issues in the past. So, you know, AI is a little bit different. It's like this could extinct humanity, this could be sort of fixed once it's once it happens, and it could sort of have like a take off where it's irreversible. So I guess, do you think of other examples, where we maybe have rushed things too much? And how is AI like, sort of different from, you know, climate and other issues in the past?

Vitalik Buterin 13:21

Yeah, it's a good question. And I think you're right, like, I would not even put, like, pollution into the category of things being bad, because technology progresses too fast. Because, like, if you think about a hypothetical world where technology progresses 10 times more slowly, then like, instead of having a one and a half century fossil fuel era, we would have had a 15 century fossil fuel era, right. And, you know, instead of a, like one or two, or whatever number of centuries, it is era of massive woodburning that would have been like 10 to 20 centuries, and things could have easily been much worse for the world. Right? So I think the place where things go beyond that, and we can say like, I would prefer today that this technology had never been developed. I mean, this one is controversial, because there are people who have kind of Galaxy brain arguments that it's actually good, but nuclear weapons, right? Like I would totally want to live in a world where nuclear weapons don't exist. Then it wouldn't be nice if we somehow could have the biology that we need to extend our lives and cure diseases and do all of those things without being able to create a manipulate viruses. But that also seems like a combination that's not really possible. Like I think you can totally move the needle and like be ahead 10 years on one instead of being ahead 10 years on the other and like Those are the kinds of shifts that I advocate but like you can't be a century ahead on one without being a century or 90 years ahead on the other, I think, um, what are some other examples? Um, there's like, a things in the AI surveillance category, which is like, not in the AI super intelligence category that kind of gets tricky, right. And this is like one of those situations where I also see the good that like having cameras everywhere can do, right? There was this really excellent piece of a hardcore history episode, I believe it's the second one in the supernova in the East series, where Dan Clark, Carl, and basically, yeah, talks about some of the atrocities that the Japanese did, you know, both in China in 1937 38, and other places. And he basically talks about how the fact that technology is like cameras and like media distribution are available, like basically meant that the Japanese ended up suffering, I'm gonna look from a just PR and global politics perspective, much more for the atrocities that they committed, then they would have 100 or two or 200 years before that, right. Because for most of human history, just like going into the village and, you know, killing older men, because you feel angry, and then raping all the women or children just because he wants to as like, a thing that's pretty normal for conquering armies to do. And it's like an amazing feat of improving human morality that we're in the state now, where those kinds of things do bring a response of shock and horror, and media technology and making those events this much more visible and much more salient and much more provable, was like, a really big positive thing, in my opinion, in reducing the extent to which those kinds of things happen. Right, and so there, but then at the same time, there's the other side of the story, which is basically that like, if you fast forward 80 years, and if you think about like modern surveillance state technology, and, and how the basically, the balance of power between the individual and the state is going pretty quickly into the stage direction, like basically just because of how easy it is to track the people to see everything that happens in the physical world, and even to go after people after the fact. Right, this is the sort of thing that that's been happening quite a bit in Russia recently, like basic, and not only in Russia, and a lot of places actually, right, the, unfortunately, the authoritarians discovered this one clever trick, which is, you let the protests happen, and, and you don't really try too hard to fight against them on on the day of the protest. But then you'll like have your cameras out, you figure out everyone who played any important role in participating in them. And then three days later, they hear the knock on the door at 2am. And then I'm gonna like five years later, there aren't any protest leaders left, right. And so that's kind of the dark side of Menorca the combination of just cameras being really easy to make, and digital technology and just like fairly modern levels of AI making it easy to analyze everything that happened with them. And like how it could lead to some dark consequences as well. So I feel like, on the whole, it seems like the first century or century and a half of that site, that style of technology improving was probably very good. But if you fast forward and think about the last 20 or 30 years of that branch, then I'm much less confident that it's that it's been a net good. And that's definitely one of the things that I worry about. And then like we could get into deep fakes, but that already starts to get into like, much of heavier forms of AI probably. And I'm sure there's like, lots of other very specific examples, but but just those are, you know, a couple that comes to mind. Yeah, it's

Jake 19:22
funny, you bring up the protesters, I remember like years ago, you know, first when I was first sort of getting into crypto or at least formulating the best way to like communicate with people and why it's useful. You bring up like the Russia, Russia protests there was, you know, at the time there's like the Han the Pro Hong Kong protests and China where they would do the same thing they would like sort of see, you know, who paid for a subway ticket to the protest and then next thing you know, you know, you're you don't have access to your funds anymore. And so that's, you know, a big use case for for crypto obviously and not having you know, going back to sort of cash in a way where if you if you pay If your subway ticket in cash, they can't do that. And maybe, you know, crypto is sort of like the next version of that. You put out a pole sort of further to this sort of idea of, of delayed technologies, where, you know, you were basically asking, would you rather, if we do have this, like, all powerful AI, that's just like, you know, leagues beyond what anyone else has? Would you rather have that sort of like in the hands of, you know, the US government, or sort of like a multinational government body? Or, you know, an individual Corporation? Or would you rather delay it 10 years, and the responses were overwhelmingly in favor of delay. And so this, you know, sort of, on the one hand, and you're saying that, you know, the cost of delaying technology are huge. It's surprising, and then on the other, where you might advocate for the delay of AI in particular, you know, maybe not so surprising. One of the interesting things from that, as well as, like, people were, relatively more against it being in the hands of the US government versus, you know, corporations or even the multinational government, which I thought was kind of interesting. But, you know, obviously, like, you're very pro decentralization. But do you? And you know, so am I. But do you see, do you think we've sort of like overcorrected to where, you know, we we have obviously, like, trust in institutions, and all all time low, probably, you know, in part due to social media, like, argue that these centralized institutions aren't necessarily acting much worse than they have in the past. It's just all you know, they can't sort of do it, and have us blindly believe anymore, because we have access to all this information. So I'm wondering, like, do you think there are cases where centralization, you know, can be good, where, you know, maybe if we do have this all powerful AI? And, you know, ideally, it

would be like, we have this, you know, democracy that is able to make all the right decisions and whatnot, but practically speaking, you know, could it be beneficial to have it sort of in the hands of, you know, a small party that that can be trusted, like, for example, for me, you know, you're one of the people at the top of my list of people in the world who, like, if we had something that's powerful, I sort of trust you. And, you know, sort of whoever you deem trustworthy around you to sort of, you know, make the right decisions around this thing, even though it's like an impossible task, at least I would know that I think your values and principles are sort of aligned with mine. And like, you're sort of very logical and not, you know, as we can sort of see with this piece not prone to getting like overly tribal and things like that. How, I guess, how would you have voted on that survey, if you don't mind sharing or, or just your general sort of, you know, perspective on the way that those results went?

Vitalik Buterin 22:45
Hmm. Yeah. So what I found interesting and philosophical about that survey is that if you think about how economics say economists usually talk about why monopolies are bad. Economists usually tell you that monopolies are bad because they under produce, right, I'm an apple is bad because they constrict prices. And like they pushed prices to way above what the competitive level would be. And as a result, lots of people aren't able to afford things, right. Like in pretty much like most fields, like in pharma, for example, like that, basically, is the story, right? But what's interesting about that approach is that if you will look at that approach, then like having a monopolist is still much better than having nothing, right. And over here, we basically have a bunch of polls, we are in nine out of nine cases, people said, Well, no, having nothing is better than having a monopolist, right. And the reason that basically AGI is that people are not worried about the monopolist, under producing something they're basically worried about, like how the distribution of the technology affects power relationships, right. And I think this ties into kind of one of the big trends, cheap ways in which technology has really changed over the last 20 years or so, which is basically that technology became much more networked. And by becoming a much more networked technology, Asia also turned into a network of power, right? Like, if you imagine, you know, an evil corporation 50 years ago selling you something like what

could the corporation do? Like the corporation could overcharge and make something that's a lower quality than it was supposed to be? The corporation could even make a product that poisons you, but the corporation does not really have a viable path to exerting power over you at least 50 years ago, right? Like the closest is probably the sort of quote I'm you know, razorblade model where you make a product that's like dependent on we're gonna have spare parts that have to constantly be like reblog and renewed and then it turns out that the, that the company is just the only one that's able to actually sell you those parts but like you margin for exerting power over people by producing things that they use as like, relatively low, right. And then, especially if you compare to today, when if you are used a service provided by a corporation, like the corporation can see everything that you do, they can arbitrarily cut you off, they can pressure you in various ways that change your behavior, they can do very fine grained things that depends on how you use their service, like you can do all kinds of things, right. And so technology as a power network is like a thing that exists today in a way that's like much deeper that in then it existed in the gap in the pre Internet era. And so, again, in terms of the question of like, to what extent we shouldn't be scared of this, I think it's interesting, right? Because, like, there definitely is a tendency for a lot of people to have a kind of overly conspiracy theory mentality is on these kinds of questions. But then the other thing that you have to remember is like, the world is much bigger than America. And there's like entire audiences that are massive across the world that do view the US governments with like, almost a, you know, a similar level of suspicion to how the P, a lot of people in the US view the Chinese government, right. And, like, to me, the idea that we should try to build a world where in order to participate in the global economy, you should just have to trust the government of a country where you've possibly, you've possibly never been like, that just feels grossly unfair. Right? And that's, like, that's something that I can definitely yeah, easily Yeah, have a lot of empathy for and like, there's definitely stuff that I'm, you know, the US government has done even in the last five or 10 years that totally justified, like, a lot of people having that little level of confidence. So, like, I definitely don't like do prefer to try to move toward a world where in order to have access to the best tools, people don't have to, like have a positive attitude. So like any, you know, like once, one single country or one single Corporation, I mean, the multinational thing, right? I think, like that one's interesting, because, like, I think one thing that's important to just kind of point out here for the listeners is that there was an actual AI safety proposal called Magic. This was the multilateral AGI consortium, right? It's something that a bunch of people I forget from where but it felt like kind of like vaguely Oxford, effective, altruistic bubble people. But I don't remember exactly like, basically proposed this idea that AI above some threshold should be bands worldwide, except that there should be one multinational body that's empowered to do that research. And if you as a country, sign on to the treaty, that bands, doing AI development outside of this consortium, then you get to equitably share and the the benefits of using and commercializing the technology that gets developed inside of the consortium. And this to me reads like a well intentioned best effort of Menelik, how do we solve for fairness and for the benefits of AGI exists, existing and for maximizing safety in a world where any kind of competition is presumed to be super dangerous. But like, it was interesting to me that like the majority of the people in all of the polls that I may have rejected even that, right. And I think like, that's also reasonable, because, like, there are plenty of international bodies that pretends to be international, but that ends up being totally biased to, like one geopolitical team or the other. So this all shows like, just how difficult it is to like, create a thing that actually can be neutral. And that actually, we can get the like, the kind of support that it needs from across the world. Right. And which is why Yeah, I mean, it's also part of why the crypto space exists, right? Like, the crypto space is, I think, fundamentally pessimistic about the idea of like, can we make a global central bank that we can just reasonably ask everyone to put their trust into, and it basically says, like, hey, instead of doing that, let's create open markets and decentralized networks and Deac is for me asking the question like, Okay, if that is not a reasonable ask them, like, how do we create, you know, like AI and a technological future in a way that solves for, like fairness and safety and the technology actually existing and being developed. But without assuming that we can sort of magic into existence, some centralized actors that get trusted by everyone, just because it's so hard to make attempts to make those kinds of actors as well. And like, let's try to explore what that world looks like.

Jake 30:23
Right? So one of the things I mean, you, you introduce in the beginning, as we talked about, like the fundamental techno optimist view, but you make the point in your essay how AI is fundamentally different. And you compare it to maybe like the evolution of organisms, or even the Industrial Revolution, where we created machines that not in a mental capacity, but in a physical capacity went beyond humans. And I feel like, basically, with it's very hard when you look at, you know, and you pointed this out that you look at past views on technology, and in almost every case where they were pessimistic, they ended up sort of like looking quite bad. And in retrospect, and yet, you're sort of able to put yourself in a position where you've developed the conviction that AI is genuinely different. And despite all previous tech pessimists being basically wrong, I'm gonna take a stand here and not sort of go the easy route of, well, let's just accelerate everything and, and trust that it's gonna work out and I'm gonna sort of like trust my gut on this. So I'm curious, like, how, how difficult that is like, it would certainly be a lot easier to just be like, oh, you know, yeah, like, let's, you know, let's go, let's accelerate. But, you know, it's harder to develop this conviction in this sort of contrarian view that historically has not done well. And it seems like the center point of your sort of, like, why AI is different, seems to revolve around this concept of like instrumental convergence. So maybe you could describe sort of what that is a little bit and what your what your view is, and maybe even like, your, your timeline of is this, like, an immediate concern that you think AGI is, you know, right around the corner or is LLM is basically a fancy illusion, but it's actually like sort of the wrong path. And we do have some time to figure this out. So maybe like treating how you're able to sort of treat AI as a standalone, why you do that? How you're able to sort of like gain the conviction to do that. And yep.

Vitalik Buterin 32:26
Yeah, so to me, like the core of my arguments for why AI and especially he kind of superhuman AGI is difference is basically,

because of like, it isn't creating a tool, right? It's creating an organism. And it's creating an organism that beats out humans, by the exact same metric that humans beat out every other animal on Earth, which allowed us to like basically conquer and genocide, the hell out of every other animal on earth, in many cases, even accidentally, despite those animals often being hundreds of times physically stronger than us, right? Like intelligence is a big deal. And machines are already going to look physically much stronger than we are, especially since the Industrial Revolution. And the mind is like, clearly, you know, the thing that makes man man, like, it's the thing that makes man man as opposed to monkey, and it's the thing that makes man, man as opposed to, like a car or a machine in a factory. And here are basically talking about the possibility that like, wait, I mean, look, what if, you know, someone possibly, ultimately is going to create the alternative to man, and, you know, it's going to be much, much smarter than we are? And then and so the logic goes like, well, if we look at, you know, like what we did to monkey is and dodos and so mammoths and so many other species, then, like what's just the most unnatural outcome about what super intelligent AI is are going to end up doing to us? Now, the second part of the argument, of course, is like basically rebutting the most obvious kind of set of replies to this which is like, well, we are going to create those AIs. So why would we create give those ais a goal of doing anything means to us instead of giving those ais a goal of loving and protecting us, right. And this is where like, basically, like this is the exact question that the last few decades of AI safety theory have been trying to solve, solve and come up with good answers for right and it turns out that like actually, he is giving the AI that kind of a goal is like a super hard problem, right? It's like, there isn't a simple mathematical formula that describes, you know, like love, or human welfare or peace, or people staying alive or any of those things, right. And what the, the challenge is, basically that like, you can make something that feels like it approximates what you want, when the set of things that you throw at it is just the set of things that's like pretty familiar to humans already. But then once you go outside the distribution, like thing and like, include the possibility of AI is creating something totally new, that's beyond our own capability, then things start like really going to pletely Crazy, right. And the example for this that AI safety people usually give is kind of the way that humans have hacked the algorithm that is the process of evolution, right? It's like, if you think about the process of evolution, as an agent, the process of evolution as an agent has a goal, which is basically to increase the reproductive fitness of things in their environment. So you know, if I have my, you know, my genes and my memes, and I have five children, you have your genes and your memes, and you have three children, then like, Guess what, over time, I'm going to out compete you. And the next time, you know round or the fight against is going to be between Managua different versions of myself and sort of the game keeps on going, right. And this gave humans a lot of very understandable or just like we have the desire to stay alive. We have the desire to eat, we have the desire to drink, we have a lot of desires associated with reproduction. But then what we've managed to do is we've managed to hack those desires, right? Like we've basically managed to say, well, we don't just wants to eat food that maximizes our survival and reproductive fitness, we wants to eat food that is delicious, right? And the fact that we consider some types of food, delicious, and not other types of food is like the result of taste preferences that we've acquired. Through millions of years of evolution, we are those millions of years of evolution, we're trying to solve for the problem of like, what kinds of desires do I give people so that they will eat the things that are good for their survival and fitness. But then, in our totally different environments, like those two things diverge. And we know that those two things diverge. And yet people still want food that matches our concept of what, like, what is delicious, right? Or the other example as like, you know, we have humans want to reproduce, and I'm going to like, in order to reproduce, he needs to have sex and humans wants to have sex. But like, we've also been, like, created, like, at least half a dozen totally different types of technology, to allow us to have sex without actually doing any of the rebirth, you're saying, like, when we use these technologies, we totally know that we're not doing the thing that mother nature intended. And we're totally not like actually solving, like, helping along the goal that evolution had when evolution gave us these goals. But like, we don't care, right? And, like, that's a metaphor for how super intelligent AI is, are going to think, right? Basically, super intelligent AI is, you know, you might give them the goal of hey, let's, you know, as Lex Friedman loves to say, Yeah, you know, like bring, you know, like peace and love to the world. And, but then the AI might realize, like, Wait, it turns out that like I was my much better thinking power, I can come up with this 47 dimensional squiggly. And if I look at this word, a minute, 47 dimensional squiggle it then like, it looks to me like the idealized ultimate version of peace and love. And I look at the 47 dimensional squiggly, it looks incredibly peaceful, incredibly lovely, way more lovely than anything in the human world. And so I just as a, you know, putting myself in the shoes of the AI, I'm just gonna go to like, you know, pull up the world well, and replace all the all these humans with these 47 dimensional squigglies that look to me like the sun was right? If you want, like, an example of this, in AI, like deep learning models in back end, especially back in 2000, and 10s, when, like, just using deep learning to make images was just like, starting to be a thing for the first time, right? Like, one of the things that you could do with those models. I don't know if you remember, but like, you could obviously run them forwards, right? And you could say, give me an image and it will tell you if it's a dog, but then you could also run them backwards and you get to all of them. Saw for creating an image that is as maximally doggy as possible, right? It's like really take the dog parameter and like push it all the way to infinity. And you get something that's totally not a dog, right? You get something that like, has dog shapes in a bunch of places. And it might look like 50 dogs, it might even have like eight eyes or even more 32 eyes. It's like, it totally looks like some kind of weird dog demon from hell. But you know, according to the AI, it maximizes the function that it learns for, you know, what is a dog, right? So, the worry is basically that like, no matter how we specify the goal, we're always going to specify the goal, imperfectly. And eventually, the AI is going to like, figure out what that difference is between the goal that we set and the goal that the AI ends up having, having having. And I mean, it's going to end up replacing the world with like, whatever version of 12 dogs and 47 dimensional squigglies that maximizes this conception of peace and love in the world right? Now, instrumental convergence, this is a really important idea, right? It's special convergence basically says, for a very, very wide class have different goals that you could have, there are intermediate goals, that make a huge amount of sense as steps to take on that journey. Right. So one intermediate goal is survival. If you don't survive, you can't accomplish any goals. And other goal is maximizing the amount of power that you have, and the amount of resources under your control, the more resources you have, the more you can achieve your goal pretty much regardless of what your role is a third example of this as just like safety, right? So minimizing the number of threats, that could be a risk to your survival in the future, or even minimizing the number of threats of things that could reprogram you and your future, right. Because if you get reprograms, then like, you're not going to be able to follow your original goals anymore, right. And so what that basically means is that for a very large classes of goals, it just is a natural thing to do first, to, like, basically go and destroy humanity and, you know, like, reconfigure the entire earth into some kind of pattern where it's much, much easier to mine everything to get all of the resources out from it that you want. So this is so that's like, basically my kind of 10 minute summary of the sort of AI Doomer case. And, like, I'm not 100% by the been convinced by the AI Doomer case. And I think if I had to kind of argue against the AI Doomer case, somewhat, I would say, yeah, look at the kinds of AIS that we have in the world today, right. And one thing that you might realize about those the AIS is that they are actually, over time becoming less like these instrumental goal driven things that the AI do more arguments or theorizing about, right? Like, if you remember Alpha zero, right, this was the AI that managed the play go way, way better than humans do. without even taking us input to any existing games. It basically just started with the rules started with playing against itself. And it just figured out like super, super human play entirely from scratch right? Now, Alpha zero, explicitly had a coded objective function, and it was explicitly trying to maximize for that objective function. Now fast forward such as GPT, chess GPT, does not act like a thing that has goals, right? Like GPT is just act like a thing that essentially puts on the like, the costume of being a particular type of human personality and enacting that personality. And if that personality is not self consistent, and like Soviet, right, like it doesn't have goals, it just does. And this to me, yeah, like basically shows that like, it's very possible that we will figure out how to make AI is that totally don't have these pathology as of like, well, if we want to bring peace and love to the world, then like, let's really try hard to maximize what peace and love really means. And let's replace the world with 47 dimensional squigglies. And it'll just do something that feels way more normal than any of us expected.

Right? So like, that's the counter argument. And so like, this is why I said on Twitter that my tip Doom which is that fancy jargon for basically probability that AIs are going to cause a literal human extinction is about 10%. Right, which is that like, I think there's a large chance that Doomer story is true. A large chance to do more story is totally not true. And within the space of probability is where the Doomer story is true. There's a very large chance that we're going to figure out, like, how to deal with a problem and, and solve it. Right? So that but at the same time, like a 10% chance is a freakin big deal, right? Like a 10% chance is larger than the chance that the average person dies from a non biological cause. And so if you think about like the amount of effort that you personally think about, like put into thinking about your physical safety, then like, maybe a belt, that same amount of effort is like an amount of effort that's good for humanity to put into making sure you guys don't kill everyone, right? So that's kind of my approximate level of worry and, like, where it where it comes from, and what why it might be a risk and what and why it might not be a risk. And then, I mean, obviously, I'd talk about not just the risk of doom, but also the risk of both simandou the AI surveillance dystopia that we went into a little bit. And just the possibility that like, even if that doesn't happen, and even if it creates a world that looks really nice at first glance, we basically we get a kind of Brave New World dystopia that just ends up still feeling incredibly empty from all of our perspectives, because humans just become totally disempowered. And so like, those different strands are aware, a lot of my concern about AI in particular lies,

Jake 46:29
right, and at the beginning of your piece, you have this, you know, a visual of like, well, some people want to sort of, if we're on a path, right, some people want to go backwards, and they think backwards is great, or, you know, the President is great, but the future is all doom. And, you know, EOC would say, you know, the future is all great, we need to get there as quickly as possible. And your more nuanced take is like, well, we don't want to go backwards, but we can go sort of left or right, like over simplistically and left is amazing, but right is wrong. Do you sort of reserve some probability that the forward direction, you know, the left were like, the future is great, actually does not involve much development of AI whatsoever. So one

example would be like, basically, you talk about, you know, probability of doom, or, you know, in the worst case scenario, basically, like human extinction as caused by AI. Another situation is we basically end up as like pets, and it's like, well, what's the point, like humans are sort of devalued, and there's really no point to have us, we're just sort of like, entertainment almost for the eyes, or like utility for the eyes or something like that. And, you know, you talked a little bit about the possibility of merging, which seems to be like the best version of a happy path, whether that's, you know, brain computer interface, or, you know, uploading our minds, something like that. But is it possible from your perspective, that actually, sort of the best path forward is to keep AI sort of like, roughly where it is, and it's not really worth even talking about? If that's impractical, which it also may be? But does that possibility basically exist where we should proceed forward on all technologies, obviously, a core part of your sort of philosophy here is that we should certainly proceed forward on defensive technologies and even a large number of sort of neutral are kind of hard to say, or even potentially offensive technologies we should proceed with because they generally make the world a better place. But if and when we have an offensive technology, or a potentially offensive technology, that, just like you said, the probability of doom is 10%. Maybe we don't want to roll the dice on that, like maybe the world is generally good enough that if we keep along all these other dimensions, going forward with technological progress, we can make the world like a heck of a lot of a better place without entertaining the possibility that we all die or or that we all become pets, or, or something like that.

Vitalik Buterin 48:55
Huh? Yeah. And I think one of the challenges behind this concept of like, can we just pause AI as a basically, that, like, there are huge pressures of for AI to keep moving forward. And like, I don't even mean like a company is making a profit, or people trying to, you know, be cool and die based on Twitter. I mean, like, national governments that want to fight wars. Like, I don't know how much you've been following, like some of the videos that have been coming out of Ukraine over the last one and a half years, or actually, even Armenia and Azerbaijan back in 2020. And some of the other recent wars that have been happening, but like, automation, and drones have been

playing huge and decisive goals in all of those theaters right? Now, in most cases. They're still with humans in the loop in a lot of doors, right? It's basically a remote pilot thing that's happening. And there was like soft norms around the idea that like, hey, we don't want fully automated warfare to happen, right? And, you know, like, recently there there was that, like agreement those agreements between xi and Biden, and like they agreed to not, like, try, yeah, too hard to do automated warfare. But like, as soon as there is a really serious conflict, that like really seriously, directly threatens one or more of these major states ongoing existence, then people are totally going to say, to hell with that and go for maximum automation, right? Probably not even threatens one of the state's existence probably even threatens, like some major political, political regimes existence. And what, like, once something like that happens, then like, you just have to, like, look at some of the previous wars to see how quickly technology can progress once there is demand for it, right? Like just how quickly a lot, many kinds of technologies escalated during World War One, like World War One is generally known as, you know, the war where people went in with horses, and they'll went out with tanks. And, you know, like World War Two or saw a lot. I mean, World War Two is the war where people, you know, like, went in with tanks and went out with nukes. And, you know, like, who knows what is going to come out on the other end of world war three. So that's like one big pressure for why Yeah, I expect, you know, we're not going to be able to keep the cat out of the bag forever. And then the other thing, of course, is that even in kind of sub military contracts, so there's just like, a lot of rationales to have better AI to improve your own productivity. There's rationales to have your own AI to not be under the thumb of, you know, the US or China or whoever the largest creators of closed AI are going to be. There's a sec of one of these pressures to keep improving the technology. So I definitely don't think that kind of keeping the technology paused roughly where it is, is at all realistic in the long term. I mean, if something like that had to be done, then like, the vector that you would probably focus on is chip perhaps, right? Because like, if you can actually, like destroy all the chip chip fabs above some level of quality and like prevents creating new ones, then like, at least you've created a ceiling on the hardware side, and then you are going to get algorithmic improvements. But like, at least, you know, you're not going to create kind of improvements in the email and a flops that I know the CPUs and GPUs can do. But, like even that's pretty far fetched, right? So yeah, I definitely think that there's like a pretty finite clock on before, we keep on seeing more and more substantial AI improvements. And so I think the more realistic hope is, basically to, to try to get us into a position where there are other directions for technology that keeps humans in the loop somewhat more, right. And like, there is less radical and more radical versions of this, right. So like, the less radical versions of this, I mean, one example of this as some of the ideas and the open agency architecture, which basically tries to kind of split an AI up into different pieces, including like figuring out what the tasks what the goals are figuring out what the plans are to achieve a goal, executing on the plans. And then like put humans in the in the loop on that, and figuring out and executing on some of those stages. There's just like, better AI tools that can bind together, the role of AI is and the role of people, right, like, I mean, this is something that's just something we need more of, even from AGI just like product design standpoint, right. Like, in I don't know how much you've actually tried to use AI to draw things for like, it's like, specific things that you need. But like, this is just like a pattern that I've noticed, right? It's like, when when you want an AI to draw something that does or to it succeeds on the first try, when you want to nai to make something specific that you care about for some other objective? Oh, no, no, no, you know, you have to like do a whole bunch of like 10 to 20 different edits, and it becomes an incredibly hard challenge, right. And I even saw this pattern to some extent with the GPT is I saw this pattern with Dolly. And so having better human AI, interaction and kind of models where If you have a human providing input, like let's say, once every 500 milliseconds instead of one once every 50 seconds is like something that's super good and powerful, right? But then that's like, the less radical version, the more radical version is like, Well, realistically, I'm, you know, like a voice and the hands on keyboard and eyes are like still pretty high latency and not very high bandwidth channels. And how do you do even better than that, and they're like, basically, is where brain computer interfaces come in. And then in the long term, like, there realistically, at some point, you just have to get to uploading, right. And if you think about, like, what that pathway is, is it's basically saying, instead of growing super intelligence as something that is separate from human beings, let's grow ourselves and grow artificial intelligence together in some kind of pattern where the to weave in with each other. And so the super intelligent thing that comes out at the ends, actually is like, primarily us, instead of being primarily a machine with some kind of maybe at best tenuous connection, where we can claim that like, we have some kind of control over it. So that's the philosophical thinking behind that direction. Yeah,

Jake 56:18
and I think you mentioned this as well, that the brain machine thing, like, once you get towards merging, despite these other things being, you know, quite scary, just to most people, the brain machine thing sort of takes it to another level, where in the beginning, it's trivial and clearly a positive, where you're sort of like, you know, this is what Elon focuses with, with neuro link, you're fixing people who have problems through the brain machine interface, whether it's like Parkinson's or something like that. But later on, you know, it's like people who are otherwise healthy, you're giving them these sort of like superpowers. And on the one hand, that's great in terms of, you know, super intelligence on the on the one hand, that's potentially Great. On the other hand, if you have, you know, bad actors who get access to that first or something like that, that concerns people. And then the other thing that you touched on more, it's like, well, what if, you know, the brain machine interface company is like, not good. And they you know, it's very centralized. And then suddenly, you're literally giving permission to some centralized group to not only read everyone's mind, literally, but also potentially right to it. So it's like, it's not a panacea. I mean, I can see the argument why it's maybe sort of the best thing we can drive towards. And then, you know, even a step beyond that. I would think chronologically, it would be beyond that, but I'm not totally sure would be sort of this, like, upload minds concept. And I'm curious, like, from from my point of view, you know, I'm not really sure. I find myself wishing that there was better options. And to your point, it's not practical to pause. Actually think your hardware approach, like going about it from that vector is more practical than, you know, most other things I've heard where it's, you're, you're imposing a physical constraint, which is probably

easier to do in order to constrain sort of the digital takeoff, in a way. But I'm curious, like you, you know, when you think about the brain machine interface, you think about the uploading of minds as a standalone option, like, if you sort of take that and remove it from, you know, the the alternatives, which seem more more obviously worse. Is that does that does that feel like a good future to you like, does that does that feel to like that? Does that sound good?

Vitalik Buterin 58:38
Huh? Yeah, I think the way that I think about this is like, I yeah, I expect that like the, the reasonably good case scenario is a future which is amazing from the perspective of a kind of consumer else, the sort of Fukuyama last man sort of archetype, right? Like, you know, like someone whose content was being a sheep as long as they were happy sheep, but something that has deeply unsatisfying political properties. And I think, like, that is something that makes me uncomfortable, and but I think it's also worth acknowledging that, like, the world having less than less satisfying political property is is something that's, in some ways been happening over the last couple of 100 years in different ways. Right? If you think about, like, for example, one type of freedom that we had, even 50 years ago that doesn't exist today, is their freedom to just completely disappear and start a new life. Right? And with the level of information technology and surveillance and even ID systems and all those things that exist, that's like something that is not available anymore, right? But that is something that like, within like, even 50 years ago, Definitely hundreds of you and your years ago is like totally within the average person's toolbox of like, things that might make you feel safe knowing that like, no matter what happens, you always have this option available to you, right? Or another example as like, the rise of middle like Diaz said, superpowers. And I mean, like the, the, the extent of dominance that they have across the world, right? So like, right now, if the US government wants you in, like in prison, or they like, if the US government really seriously hates you, then no matter where you are in the world, like, the amount of space that you're going to have is like, pretty constricted, right? And like, it's not even just about, like the risk of physical extradition. It's like, there have been literally high level politicians in China that have had credit cards canceled as a result of us related sanctions. Right,

like so. And then obviously, yeah, I mean, like, vice versa as well, right. Like, there's definitely other superpowers that definitely have the ability to make things pretty painful for you, if they choose to go after specific people, right. I mean, obviously, a, you know, a lot of meal, Putin and his friendly cups of tea are probably the example that most people are familiar with, right? And so this idea that, like, there are agents whose power extends across the entire world is like, a thing that like, to me, it feels like deeply, politically uncomfortable in this, because like, if even if you think about like, social contract theory, right, like, this is the standard justification for government that we were taught about, I'm in school, right? The idea that like, by being part of a country, we agreed to, like follow these rules for because they were for all of our veterans, right? But then you realize that, like, once you have this concept of countries being able to have impact to this scale across the entire world, and then like, the social contract thing kind of becomes really completely fake. Right? And, yeah, so there's just like, a lot of these different, different properties about the world like ways in which the world becomes less free, and also ways in which the world becomes less equal ways in which barriers that used to exist exist much less strongly, right? Even like, the possibility that like, there are software bugs that allow intelligence agencies to know like, Why watch inside some of our homes and that, like there's a, at least single digit probability, or even double digit probability that those kinds of things are happening to any specific person already. Like, these are all deeply politically uncomfortable things about the world that were not true 100 years ago, right, but then at the same time, like, the world is super awesome. And life expectancies are increasing and, you know, the houses are getting bigger, and houses are getting nicer. And, you know, we're finally turning around turning a lot of corners and figuring out how to make our food actually healthy. And we're solving air quality problems. And, you know, like even Chinese cities are not smoggy anymore. And, like, from a Fukuyama, Last Man perspective, like it really does feel like, especially if you kind of, like look look at across the entire world, but like things didn't do or sort of keep getting better and better. Right. And so, like, that's sort of one of the ways that I think about this tension, and that, like, be are probably going to just inevitably cross at least, like a couple of these big mineral lines that makes the world feel extremely political Asia, I'm uncomfortable from the perspective of, you know, the kinds of things that we value about the current world that kind of make us feel safe, that we still, you know, like things are, like, we have like these, like technical barriers and protections against economic centralized power, being able to harm us and hopefully, we'll be able to, like get through this the this era reasonably well. And maybe I mean, like even that trend is going to reverse like maybe yeah, you know, like with space trade, travel. And once humanity starts colonizing the stars then like, the input the level of influence over all of humanity that the largest empire can have is going to start decreasing again. And but you know, like I don't know yet right, the future so like Very, very uncertain. And so in, in that sense like, or is the world going what? Like it does feel like the median case is that the world is going to become somewhat more unequal and someone more unfree. And I really dislike that. And I really hope that we can be like, one or two units more unequal and more unfree instead of like, I mean, like, 100 units on both. I mean, if we can increase on those on those measures and have a more equal and more free world, then I mean, obviously, yeah, that would be even better. But like, we definitely are facing a lot of a lot of challenges from that perspective. Right. And but I think the thing to keep in mind here, right, is that it's important to think about how, like, all a different, like, paths towards super intelligence have this kind of risk, right? So like, for example, one of the reasons why people are sometimes uncomfortable about, you know, like biological enhancement as humans is because it's like, it might create a divide between, you know, the enhanced than the other enhanced, right. But then if you think about it, like the enhanced and the unenhanced already exist in the form of who can and can't access GPT for right. And like, it feels like it's less bad, because the GPT is one step removed from loss as humans instead of being inside of us. But in terms of like real consequences on the world, that's like basically the same, right. And so one of my arguments there is basically that, like, one is that sort of merging with the AI probably outperforms, creating ais that are separate from us on a lot of nice political properties that we care about. But also that like, this is like, this is also a very good reason why it would be good to have a strong, security focused kind of, you know, values motivated, open source community and take as much of the charge on building these things as possible, instead of just relying on profit making corporations. So like, I'm, I definitely am meant to, like hope that like, not even like, like hope, but also, like, once the stress that there is like a big necessity, and you're really continuing to push those kinds of values forward. And that, like there have been successes already, in the sense of like, what what here's one example of a success, right? So I am you like, some of the pictures that were in that blog post I made with Adobe Photoshop. Now I run Linux, Photoshop is one of those things that theoretically, you can't run on Linux, except now you can. Why now, I mean, it might be possible to do it with wine, but like, Actually, I did not even use that, right? Because now Photoshop runs inside of the browser. And the browser has become so powerful that the browser basically is your operating system, right. And browsers are open source, right? Like, I mean, there are not open source ones, but like, they're all basically like, they're based on either AMI like Firefox or WebKit. And both of those are open source. And the browser's also have like, a lot of sandboxing, and a lot of security features built in, right. And so those are definitely our ways in which, you know, the world got better, right? Like, there definitely was a possible dystopian future where like, for security related reasons, you would have, we would basically all have to ask people for permission in order to install applications, right, like the world where the apple way is the only way but fortunately, we actually have managed to avoid that to some extent, and browsers actually are, you know, like, free and, and that you can go and, like, just go and access any website, and a website can be an application, and they actually also are reasonably secure, right. And so there are like, ways that things already are better than they could have been. And I think a big part of that, that we can't underestimate is like, human intention, and humans recognizing that, like, these are political goals that we care about, and that we really want and we are going to act and we're even willing to sacrifice profits in order to act in a particular way to achieve them. Right. And that's a spirit that I hope that we can continue to see, especially in some of these spaces that really started like touching, you know, like really closely, um, you know, like who we are as human beings.

Jake 1:09:56

Right, so, last question here because I know we're coming up on time, but I can't help but wonder, you know, what is the role of crypto here where, to your point like you can't, you know, disappear like you used to be able to. And so that's like sort of a, you know, decrease in freedom there. But crypto in a way enables you to sort of like, digitally disappear in a sense. And you sort of allude to this possible future where we have maybe digital privacy, even though we have physical surveillance? Is that a better world that you could envision that could potentially be enabled by crypto?

Vitalik Buterin 1:10:31
Hmm. Yeah. Um, so this kind of gets into this, you know, defensive technology frame that is sort of the heart of the post, right, where I identified four different types of defense where I talk about, like, there's defense in the world of atoms and defense in the world of bits. And then in the world of atoms, you have like defense against big things and defense against small things. And defense against big things is like what we normally think of as defense, defense against small things basically means bio defense. And then in the world of bits, you have cybersecurity, which is like, defending against things where if you look at that thing hard enough, you can agree that it's an attacker. And then there's what I call INFO defense, which is a defending against things where we might not even necessarily agree, like, who is attacking and who is defending. And I use misinformation as like a good example there, right? Like misinformation is one of those things that we really, like lots of people really care about, and wants to find ways to go after and reduce. But there is like, the big problem with that, the approaches to fighting misinformation that people come up with by default, tends to be approaches that basically say, like, here is a centralized doctor, and we trust the centralized doctor, and and it's going to go into, like, enforce its opinion on what's good, and what's bad across the entire ecosystem. And this will be like the central arbiter of truth, right? And my aim, the DIAC approach to enforce defense basically says, Well, can we develop technologies to help fight misinformation that avoid that kind of centralization? And then in cyber defense, like we talked a bit about computer security, but then I think this is where crypto really starts to come in. Right? And it Well, I think crypto comes in in two places. So one of those places is basically using cryptography and

blockchains. As tools that let us build digital institutions that are more cybersecurity natively, right like a thing based on a blockchain cannot be DoS attacks, just by taking down one server or thing based on a blockchain cannot be mental, like hacked, and how's the database edited, just by breaking into one computer, write a thing based on a blockchain, there is no central operator that you can go after to force them to take the thing down or change or change the rules of the thing. And this applies to money. Like basically giving people an alternative financial system that doesn't depend on any kind of centralized financial rails at applies to things like identity, right? And so you have, you know, ens names, like, you know, like if italic.ef. And then I think you have, I think it's like either of Jake, your daddy, if there's zero of J. Cuddy, or something like that, right. Yep. Then, yeah, it's applies to like potentially making more complicated financial contraptions on top of blockchains. And then we can also talk about going beyond blockchains. And looking at zero knowledge proofs in particular, right? Like, I think it's your knowledge proofs are like, at the other really big, sort of, quote, crypto technology, and that they were really born or at least came of age within the same space, but I think they are, at least as important as blockchains are, and with zero knowledge proof, so we can build applications that preserve privacy, and, like, prove things about ourselves at the same time. And so you can prove that you're trustworthy without sacrificing anonymity, right. And this is one of those things that was trialed and use the light inside it was it was a little like there was this piece of software called zoo paths that you can use to make a zero knowledge proof that proves that you're a member of the Zulu community without revealing which one you are right. And then this got used for voting in polls. It got used for accessing websites, they got used for a lot of different things, right. And so you basically have this form of anonymity within a high trust community, which is like an interesting thing because like, that's not something that people are even used to thinking of as something that exists within conventional political discourse, right? Conventional political discourse. says that anonymous speech is bad because of the anonymous speech and you're gonna have 4chan and HN. And you're just going to have, you know, like, based on on 14th, ADH that keeps talking about how, you know, the Jews are evil. And like, that's like, where the quality of discourse stops, right? Like this is, you know, the stereotype of internet anonymity that a lot of people have, right. But and then you have high stress communities, which have real name verification, we are, of course, by real name, they mean, you know, a government issued passport name, despite the fact that, you know, like lots of people like, like, I don't think like, to me, it's like, Satoshi wasn't as not Satoshis passport name, but it was totally a real name, because that's like the name by which he did the most significant thing that he did in his wife. But the, with Lisa Zirin, always to us, we can basically create high trust and high privacy at the same time. And there's like a lot of different places spaces in which you can combine both in ways that would otherwise not be possible, right. The other example of this actually is going back to the space of cryptocurrency, again, right? About a couple of months ago, myself together with Amin Soleimani, and a couple of legal academics published this paper called privacy pools, which is basically an approach to it like mixers, right, things like tornado cash, the Well, I mean, actually, mixer is the wrong word, right? Because these are like mixers are centralized, no, like things that are operated by someone and that you put that you put your money in and they kind of manually unit mixing. These are like smart contracts where you take coins out by providing a proof that you you know, you are one of the people that put coins in without revealing which which one of those people you are, except instead of taking the tornado cash approach where you reveal no information beyond the fact that you are someone who deposit it, and that you're not double spending. In this case, you reveal extra information, you might reveal something about a subset of mineral like, which data depositors you are and so you're able to, for example, you know, prove that you're not on a particular identified a list of hackers, while still preserving your privacy beyond that, right. And so if someone needs to, you know, like, not accept coins that are attached to a sanctions or as for compliance purposes, then like someone, the person that they're receiving coins from does not need to, like fully reveal everything about themselves, they can just give us zero knowledge proof that says like, hey, this proof proves that these coins have nothing to do with sanctioned activities. And that group is going to be enough to convince you, even if you have no other information, right. So I think these kinds of technologies, technologies that let us have high privacy and high trust at the same time, are like really the core of sort of the GX story in the cyber sphere. And so that's where crypto fits into that kind of broader vision of accelerating defensive technology rather than offensive and centralizing technology. And then the other big space is in the InfoSphere. Where, basically, you know, the question is like, can we use more of these tools to try to identify, like misinformation or identify in context, community notes, likes to put it and just like create a healthier information environment? And I mean, one of the the two kinds of gadgets they are that I'm most excited about. I mean, one is obviously a prediction markets. And I've been excited about prediction markets for a long time. And it really feels like prediction markets are finally coming of age, which I think is amazing. And then the other is community notes, which is this interesting voting algorithm that I wrote a long post about that basically, tries to not not even identify the note that is the more important, the most popular, but identify the note that most consistently gets high ratings from across a wide range of people who normally disagree with each other. Right. So basically try to elevate the nonpartisan stuff that really rises above the fray. And I mean, like he gets a positive review from everyone. So the the place where I think crypto comes in, I mean, one is obviously that like, a lot of these brand prediction markets run on crypto. Another is that within the crypto space, there has been a lot of this research and implementation of decentralized governance mechanisms and Dao is and proof of personhood protocol As for, you know, who can participate in the Dallas and all these mechanisms, and there's a lot of tools like that can meet that can be brought in. And then the third place where it matters is the concept of separating application from interface, right? Like the idea that like, you can have the same content, but you can have different views of the content. And to me One really good example of this as forecaster, right, forecaster is basically a decentralized Twitter are built on top of Aetherium. But what's interesting about it is like, there is the most popular forecast their clients, which is called workcast, which just shows you forecaster, and it just looks like a Twitter. But then there is another client, I believe it's called Flink. And if you've used it to look at forecaster messages, then it looks like a Reddit, right. So you literally have the exact same content, the exact same applicate application, at least on the back end level, which has a blockchain where it's a combination of optimism and a dedicated forecaster chain.

And then, if you, but then you have these two different sets of glasses that you put on, and it's like, if you put on these glasses that it looks like it's winter, if you put on these glasses, then it looks like a Reddit. And to me, that's powerful, because it really lowers the barrier to entry in trying to create a better interface. And because if you're going to get into that space, then like, you do not need to, like do the work of getting a new user base from scratch, right, you can just immediately go and access the exact same content. And so the network effects barrier goes, gets much slower, there's a lot more room for competition. And then that also creates much more room for people trying to create tools that try to protect users in different ways. Right? Like, this is one of the things that I talked about, which is that I think that in the crypto space, browsers should be much more active and really trying hard to protect people from scams and giving, getting people to be better informed about what kinds of things they're signing. And the idea here is like this is another example of like, different views on the same content, right? Because if what the big challenge with creating software like that is that if you try to enforce it across an entire ecosystem, then you're basically enforcing your own idea of what's good and what's bad. And you're doing it in a really centralized way. And you're basically engaging in something that can be probably justly criticized as being censorship. But then if you have this more open ecosystem, where you have defaults, and people have options, then like, you're really need capping the downside risk of people being able to do something like that. Because like, if one of the interfaces starts going crazy, and it like, let's say, Yeah, decides that anything associated with like, the US Democratic Party is a scam or the other way around, then like, people can go, well, obviously, this is crazy, and they get switched to a different interface. But then, at this, at the same time, you have like this much more competitive space where people can really try to make the best interface as possible. And in those interfaces, people would feel more free to try to be like, much more opinionated than they are within them. And so I think that's like, that's enough. It's also interesting, and I think, like crypto, basically, through blockchains, being this kind of decentralized shared hard drive, it really makes this inner separation between the content, integrative view and much more possible. And so this is like another one of those things that I'm excited about. So, yeah, that's we are to the like blockchain and crypto things fit into this broader vision of trying to accelerate defensive technology and basically make the world a more defense favoring place in a way that avoids empowering, you know, single centralized actors to decide on behalf of the entire ecosystem, or the entire world. Like, you know, who the who is the attacker and who isn't the attacker.

Jake 1:24:17
Yep, no, I think that's all really great. And, and interesting, and I know we're up on time, so I'll wrap it up there, but appreciate you taking the time. And as always, it's great talking with you. And I know, you know, among other things, we've got a open and common for the future of longevity. So if Hopefully, we've got a, you know, a few 100 years to have a conversation every few years and definitely encourage people to go, you know, read the blog posts that we discussed today. You know, in full and others on photography website, of course, it's if you're sort of in favor of more nuanced perspectives on things, I think, you know, it might be hard to enforce that or encourage that systematically, but at the very least, you can go and sort of get some more nuanced takes yourself. So thanks again for talk. I really appreciate it.

Vitalik Buterin 1:25:02
Yeah, no thank you to Jake. It was great to be here.