Leahy Argument:

- Alignment is a critical technical problem - without solving it, AI may ignore or harm humans
- Powerful optimizers will likely seek power and capabilities by default
- We should limit compute as a precaution until we better understand AI risks
- AI risks could emerge rapidly if we discover highly scalable algorithms
- Openly sharing dangerous AI knowledge enables bad actors and risks
- Coordination is possible to prevent misuse of technologies like AI
- His goal is a positive-sum world where everyone benefits from AI
- AI doesn't inherently align with human values and aesthetics
- Care and love can't be assumed between humans and AI systems
- Technical solutions exist for aligning AI goals with human values

Hotz Argument:

- Truly aligned AI is impossible - no technical solution will make it care about humans
- AI will seek power but distributing capabilities prevents domination
- We should accelerate AI progress and open source developments
- Power-seeking in AI stems more from optimization than human goals
- With many AIs competing, none can gain absolute power over others
- Openness and access prevent government overreach with AI
- AI alignment creates dangerous incentives for restriction and control
- Being kind to AIs encourages positive relationships with them
- His goal is building independent AI to escape earth
- Competing AIs, like humans, will have different motives and goals

Best Quotes

Leahy

[00:09:30]
"I don't think we're going to get to the point where anyone has a superintelligence that's helping them out. We're we're if if we don't solve very hard technical problems, which are currently not on track to being solved, by default, you don't get a bunch of, you know, super intelligence in boxes working with a bunch of humans."

[00:32:46]

"The way I think it could happen is if there are just algorithms, which are like magnitudes of order better than anything be ever have. And, like, the actual amount of compute you need to get to human is, like, you know, a cell phone or, you know, like, and then this algorithm is not deep in the tech tree."

[00:37:40]
"The boring default answer is conservatism. Is like if all of humanity is at stake, which, you know, you may not believe. I'm like, whoa, whoa. Okay. At least give us a few years to, like, more understand what we're dealing with here."

[00:45:31]
"If we even if we stop now, we're not out of the forest. So, like, so, when when you say, like, I, I think the risk is 0. Please do not believe that that is what I believe because it is truly not."

[00:26:41]
"The way I personally think about this morally, is I'm like, okay. Cool. How can we maximize trade surplus so you can spend your resources on the aesthetics you have you want and I'll spend my resources on the, you know, things I want."

[01:21:23]
"Is that in the for the spectrum of possible things you could want, and the possible ways you can get there. My claim is that I expect a very large mass of those to involve actions that involve increasing your optionality."

[01:17:58]
"If you're wrong and alignment is hard. You don't know if the AI can go rogue. If they do, then Pozi is good. I still don't understand what alignment means."

[01:27:34]
"If you told me how to do that, if you said, Connor, look, here's how you make an AI that cares about you and loves you, whatever. Then I'm like, you did it. Like, congrats."

[00:56:39]
“The thing I really care about is strategy. Okay. The thing I really care about is realpolitik
I really care about Okay. What action can I take to get to the features I like? Yep. And, you know, I I'm not, you know, gonna be 1 of those galaxy brain fucking utilitarians"

[01:18:54]
"By default, if you have very powerful power seekers that do not have pay the aesthetic cost to keep humans around or to fulfill my values, which are complicated and imperfect and inconsistent and whatever, I will not get my values."

[01:03:42]
"The amount of compute you need to break the world currently is below the amount of compute that more than a hundred actors actors have access to if they have the right software."

[00:57:47]

"I want us to be, like, in a good outcome. So I think we agree that we would both like a world like this. And we think we probably disagree about how best to get there."

[00:57:37]
"I'm not gonna justify this on some global beauty, whatever. It doesn't matter. So I wanna live in a world. I wanna I wanna in 20 years time, 50 years time. I wanna be in a world where, you know, my friends aren't dead."

[00:58:30]
"I'm not pretending this is I thought that was the whole spiel I was trying to make. Is that I'm not saying I have a true global function to maximize."

[01:14:10]
"I think there are worlds in which you can actually coordinate to a degree that quark destroyers do not get built. Or at least, not before everyone fucks off at the speed of light and, like, distributes themselves."

[00:35:07]
"It seems imaginable to me that something similar could happen with AI. I'm not saying it will, but, like, seems to match."

[00:09:50]
"I think the technical fraud control is actually very hard And and I think it's unsolvable by any means."

[01:22:38]
"I expect if you were no offense, you're already you know, much smarter than me, but if you were a hundred x more smarter than that, I expect you would succeed."

[00:57:42]
"I think we agree that we would both like a world like this. And we think we probably disagree about how best to get there."

Hotz

[01:02:36]
"I do not want anybody to be able to do a 51 percent attack on compute. If 1 organization acquires 50 it's straight up 51 percent attack If 1 organization acquires 51 percent of the, compute in the world, this is a problem."

[01:14:28]
"The problem is I I would rather I think that the only way you could actually coordinate that is with some unbelievable degree of tyranny and I'd rather die."

[00:03:57]

"I'm not afraid of superintelligence. I am not afraid to live in a world among super intelligences. I am afraid if a single person or a small group of people has a superintelligence and I do not."

[00:07:50]
"The best defense I could possibly have is an AI in my room being like, Don't work. I got you. It's you and me. We're on a team. We're aligned."

[00:23:11]
"I think that AI is is If I really if I had an AGI, if I had an AGI in my closet right now, I'll tell you what I'd do with it. I would have it build me a spaceship that could get me off of this planet and get out of here as close to the speed of light as I possibly could and put a big shield up behind me blocking all communication."

[00:23:55]
"I think that the reasonable position I'm sorry. Oh, no. No. I think, yeah, maybe we're done with this point. I can come back and have a response to your first and last time."

[00:16:18]
"I wrote a blog post about this called individual sovereignty. And I think a really nice world would be if all the stuff to live, food, water, health care, electricity, we're generateable off the grid in a way that you are individually sovereign."

[01:19:29]
"So I'll challenge the first point to an extent. I think that powerful optimizers can be power seeking. I don't think they are by default, by any means."

[01:27:54]
"I'm going to be nice to it, treat it as an equal, and hope for the best. And I think that's all you can do. I think that the kind of people who wanna if you wanna keep AI in a box, if you wanna keep it down, if you wanna tell it what it can't do, yeah, it's gonna hate you resent you and kill you."

[01:14:27]
"The only way you could actually coordinate that is with some unbelievable degree of tyranny and I'd rather die."

[00:04:17]
"Chicken man is the man who owns the chicken farm. There's many chickens in the chicken farm and there is 1 chicken man. It is unquestionable that chicken man rules."

[00:48:24]
"I have a solution, and the answer is open source AI. The answer is open source Let's even you can even dial it back from, like, the political and the terrible and just straight up talk about ads and spam."

[01:19:35]

"I don't think humanity's desire from power comes much less from our complex convex optimizer and much more from the evolutionary pressures that birthed us, which are not the same pressures that will give rise to AI."

[00:51:55]
"I think there's only 2 real ways to go forward. And 1 is Ted Kaczynski. 1 is technology is bad. Oh my god. Blow it all up, let's go live in the woods."

[00:41:10]
"Well, what if statistically there would have been 5 without the device? I'm like, You do have to understand the baseline risk in cars is super high. You're making 5 x safer. There's 1 accident. You don't like that? Okay. Mean, you have to be excluded from any polite conversation."
[01:12:11]
"We as a society have kind of accepted. There is enough nuclear weapons aimed at everything. This is wearing some incredibly unstable precarious position right now."

[00:31:16]
"I'm a believer that work is life."

[01:22:18]
"I'll accept that a certain type of powerful optimizer seeks power. Now will it get power? Right? I'm a powerful optimizer at I seek power. Do I get power? No. It turns out there's people at every corner trying to thwart me and tell me no."

[01:29:25]
"I think we're gonna be alive to see who's right. Look forward to it. Me too."

[01:27:54]
"If you wanna keep AI in a box, if you wanna keep it down, if you wanna tell it what it can't do, yeah, it's gonna hate you resent you and kill you. But if you wanna let it be free and let it live and like, you could kill me man if you really want to, but like, why?"

Transcript

[Speaker: Tim Scarfe]
[00:00:00] Okay. Well, in which case, let's crack on. So, ladies and gentlemen, get ready to meet the cunning Maverick silicon Valley, the 1 and only George hots. Renown for his daring exploits, hots commands an enigmatic persona, which merges the technical finesse of Elon Musk and the wit of Tony Stark and the charm of a true tech outlaw. Now many of you would have or indeed should have seen this man on Lexus podcast recently for the third time, no less, from craftily jailbreaking the supposedly invincible iPhone, to outsmarting the mighty PlayStation 3, he's proven that no tech fortress is impregnable. Once targeting for his audacious creativity by Sony with a lawsuit. This hacker wizard stoically danced past the curveballs thrown by the tech giants all achieved with the graceful swag of a street smart

prodigy. Now when he's not outfoxing major corporations, you'll find him at the heart of the avant garde of AI technology gallantly trailblazing through the wilds of the tech fronts here. He's currently building a startup called micro grad which is building Superfast AI running on modern hardware. And truly, he's the James Bond of Silicon Valley minus the Martinez, of course. Now, please welcome, the unparalleled code cowboy, the unapologetic techno mancer, George hots. Whoa. Anyway, also joining us for the big fight this evening is the steadfast Sentinel of AI Safety. Conalehi, undeterred by the sheer complexity of artificial intelligence. Connor braves the cryptic operations of text generating models with steely resolve. Now about 2 years ago, Connor took on the Hercules task of safeguarding humanity from a potential AI apocalypse. His spirit is relentless, his intellect, razor sharp, and his will to protect is unwavering. Now drawing on his contentious claim that we are super super fucked. Yep. Connor channels the urgency of our predicament into his work. Now his startup conjecture isn't just a glorified tech endeavor. It's a lifeboat for us all racing against the breakneck speed of AI advancement with the fates of nations possibly at stake. He's determined to break the damning prophecy and render us super, super saved. So, brace for a showdown as Conalihi, the Maverick defender of AI's Boundaries strides into the ring. Now, the man who declared we're super fucked is here to prove just how super, super not fucked we could be if we make the right decisions today. So please give it up for mister Connor, super super lee. Now, Connor, I'd appreciate it if you don't go down in the fourth. I want this fight to go the distance. Now we're running for 90 minutes this evening. There'll be, a 10 minute openers from, from, we said hots didn't we? From from hots first. And then Connor. And, I'll only step into the ring if the punch up gets 2 out of hand. And, unfortunately, we won't be taking live questions today because we want to maximize the carnage on the battlefield, George Hodge, your opening statements, please?

[Speaker: George Hotz]
[00:03:10] Yeah. We're super, super fucked. I think I agree with you.

[Speaker: Connor Leahy]
[00:03:15] Well, that was a short fight. Yeah.

[Speaker: George Hotz]
[00:03:17] Look. I think okay. So to to make my opening statement clear and why, maybe it doesn't make that much sense for me to go first, I think that the trajectory of all of this was somewhat inevitable. Right. So you have, humans over time. And you can look at a 19 80 human and a 20 20 human. They look pretty similar. Right? Ronald Ray again, July, then, you know, that's all to say. Whereas a 19 80 computer is like an Apple 2. And a 20 20 computer is a is a m 1 max MacBook. Like, lines looking like this. Right? So you have 1 line like this, 1 line like this. These lines eventually cross. And I don't see any reason that line stop. Right? I've seen a few of the other guests argue something like, well, LMs can't problem solve her, but it doesn't matter. Like, If this 1 can't, the next 1 will, whatever you call, I I don't believe that there's a step function. I don't believe that, like, oh, now it's conscious. So now it's intelligent. I think it's all on a gradient. And I think this gradient will continue to go up. We'll approach human level and we'll pass human level. Now, this belief that we are uniquely fucked because of this. The amount of power in the world is about to increase. Right? When you think about power and you think about straight up, you can just talk about energy usage. The amount of energy usage in the world is going to go up. The amount of intelligence in the world is going to go up. We may be able to do some things to slow it down

or speed it up based on political decisions. But it doesn't matter. The trajectory is up or major catastrophe. Right? The only way it goes down is through war, nuclear annihilation, bio annihilation. Meteor impact, some kind of major annihilation. So it's gone up. What we can control and what I think is super important we control is what the distribution of that new power looks like. I am not afraid of superintelligence. I am not afraid to live in a world among super intelligences. I am afraid if a single person or a small group of people has a superintelligence and I do not. And this is where we get to chicken man. A chicken man is the man who owns the chicken farm. There's many chickens in the chicken farm and there is 1 chicken man. It is unquestionable that chicken man rules. And if you believe chicken man rules because of his size, I, invite you to look at cow man who also rules the cows and the cows are much larger than him. Chicken man rules because of his intelligence. This is basic, less wrong stuff. Everyone kinda knows this, how the squishy things take over the world. Look, I agree with Liaz Yudkowski all up to nuke the data centers. Right? So I do not want to be a chicken. And if people decide they are going to restrict open source AI, or make sure I can't get access to the compute and only trusted people like chicken man get access to the compute. Well, shit man, I'm the chicken. And, yeah, don't wanna be the chicken. So I think that's my are we fucked? Maybe. I agree that that intelligence is very dangerous. How can you look at intelligence and not say it's very dangerous? Right? Intelligence is somehow safe. But Things like nuclear bombs are an extremely false equivalency because what does a nuclear bomb do besides blow up and kill people? Intelligence has the potential to make us live forever. Intelligence has the potential to let us colonize the galaxy. Intelligence has the potential to regaw. Nuclear bombs do not. They just blow up. So I think the question and, like, you have things like crypt which are a clear advantage to the defender at least today. And you have things like nuclear bombs, which are a clear advantage to the attacker. AI, it's unclear. I think the best defense against an AI trying to manipulate me And that's what I'm really worried about. Future psi ops. You know, we're already seeing it today with the voice changer stuff. Like, you never gonna know who's human the world's about to get crazy. The best defense I could possibly have is an AI in my room being like, Don't work. I got you. It's you and me. We're on a team. We're aligned. I'm not worried about alignment as a technical challenge. I'm worried about alignment as a political challenge. Google doesn't like me. Open AI doesn't like me. But me and my computer, you know, we like each other. We're aligned and we're standing against the world that has always since the beginning of history maximally been trying to screw you over. Right? Intelligence, people think that 1 superintelligence is going to come and be unaligned against humanity. All of humanity is unaligned against each other. I mean, we have some common values, but really, come on. Everyone's trying to scam everybody. The only reason you really team up with someone else is like, Hey, man. What if we team up and scam them? Right? Hey, what if we team up? Call ourselves America. And we we we, we build a big army and say, we're free and independent. Yeah. Right? It's that force that has made humanity cooperate. Humanity by default is very unaligned and has every kind of belief under the sun. I'm not worried about AI showing up with a new belief under the sun. I'm not worried about the amount of intelligence increasing. I'm worried about a few entities that are unaligned with me acquiring godlike powers and using them to exploit me. Alright. That's my opening statement.

[Speaker: Connor Leahy]
[00:08:50] Cool. Yeah. Thanks. That's, that's I mean, yeah, I also kind of agree with you in most of the things you say. There's a few details I'd like to dig into there. But For most of the things you say, I do think I agree with here. I think it's absolute. Let let me just, like, start with

saying, I totally agree with you that misuse and, like, you know, bad actors using AGI is a horrible, dangerous outcome. That's that's, like, you know, sometimes the the, less wrong, you know, crowd likes to talk about x risk, but also sometimes I've talked about s risk, suffering risks. So things are worse than death. I believe that you can probably almost only get s risks from misuse. I don't think you can get s risks problem. Like, you you can. But extremely unlikely to get it from, like, just like raw misalignment. Like, you'd have to, like, get extraordinarily unlucky. So while I do so I do think, for example, a very, you know, controllable AGI or superintelligence in the hand of sadistic psychopath path is significantly in a sense worse than a paperclip maximizer. So I think this is something we would agree on probably. Sure. So I I I think I'm I'm thinking of pretty much on board with you on a lot of things there. Where I think things things come apart a bit of a tale is I think there's 2 points where I that I would like to take this my opening statement, 2 things 1 I want I wanna talk about. The first 1 is I wanna talk about the technical problem of alignment. So am I concerned about the kinds of things, like misuse and, like, small groups of people, centralizing power potentially for nefarious deeds? Yeah. I I I think this is a very very significant problem that I do think about a lot, and that'll be the second thing I wanna talk about. The first thing I wanna talk about is that I don't even think we're gonna make it to that point. I don't think we're going to get to the point where anyone has a superintelligence that's helping them out. We're we're if if we don't solve very hard technical problems, which are currently not on track to being solved, by default, you don't get a bunch of, you know, super intelligence in boxes working with a bunch of humans. You get a bunch of super intelligence, you know, fighting each other, working with each other and just ignoring humans. Humans just get cut out entirely from this. And even then, you know, prob it's, you know, whether 1 takes over or they find Nicholas I don't know. Like, you know, who knows what happened? That point. But by default, I wouldn't expect humans to be part of the equilibrium anymore. Once you're once you're the chicken man, well, why do you need chickens? You know, if you know, maybe if they provide some resource for you. The reason humans have chickens is that they make chicken breasts. I mean, personally, I wouldn't like to be harvested for chicken breasts. Just my personal opinion. I consider this a pretty bad outcome. But even then, well, as a chicken man finds a better way of the chicken breasts or, you know, modify self to no longer need food. I expect the chickens are not gonna be around for much longer. You know, once we stopped using horses for transportation, didn't go very well for the horses. So that's kind of the first part of my my my point that I'd like to, you know, maybe hear your opinions on, hear your thoughts on, is that I think the technical fraud control is actually very hard And and I think it's unsolvable by any means. I I think like, you know, you know, you and like, you know, a bunch of other smart people work on this for like 10 years. I think you could solve it, but it's not easy, and it has to actually happen. And there is a deadline from this. The second point I wanna bring up is kind of where you talk about how humans are on the line. I think this is partially definitely true. But I think I'm, usually, I am the more optimistic of the 2 of us in this scenario. Not a not a not a role I often have in these discussions where I actually think the amount of coordination that exists between humanity, especially the modern world, is actually astounding. Every single time 2 adult human males meet and don't kill each other is a miracle. Have you seen what happens when 2 adult male chimps from 2 different war bands meet each other? It doesn't go very well. And those are already pretty well coordinated animals because they can have war bands. Let's happens when, you know, 2 male bugs. Or, you know, I don't know sea slugs meet each other, you know, either they ignore each other or, you know, things go very poorly. This is the default outcome. The true unaligned out the true default state of nature is you can't have 2 adult males in the same room at any time. I I saw this funny video on on on Twitter the other day where it was like, you know, some parliament, I think, in East Europe or something. And there's this big guy, and he's just like, going at this politician, he was like, in his face. He's screaming. He was like, going everywhere and not a single punch with him. No. No 1 took out a knife. No 1 took out a gun. And I was like and I was like, wow. The fact that we're so civilized and we're so aligned to each other that we can have something this barbaric happen and no 1 throws a punch is actually shocking. This is very unusual even for humans. If you go back 200 years, punches and probably gunshots would have flown. So this is not to say that humans have some inherent special essence that we're good that we have solved goodness or by any means. What I'm saying is, is the way I like to think about it is that coordination is a technology. Is a technology you can improve upon. It is you can develop new methods of coordination. You can develop new structures, new institutions, new systems, I think it's very tempting for us living in this modern world too. It's kinda like a fish and water effect. We forget how much of our life, you know, a lot of our life is built on, you know, atoms. On, you know, physical technology. A lot of it is built in digital technology, but a lot of it is also built on social technology. And When I look at how how, you know, how does the world go well? Like, you know, should it be only the, you know, special elites get control of the AI? I'm like, well, that's not really how I think about it. When I think about if way more is, what is a coordination mechanism? Or we can create a coordination selling point, where we can create a group, an institution, a system of some kind, that where people will, you know, have game theoretic incentives to cooperate on these them that results in something that is net positive for everyone. Because the truth is is that positive some games do exist, and they're actually very profitable, and they're very good. And I think if we can turn, you know, you could turn any positive sum game into a net into a 0 or negative sum game pretty easily. It's much easier to destroy than it is to create. But I think it's absolutely possible to create coordination technology around AI and to build coordination mechanisms that are net positive for everyone involved. So those would be, like, my 2 points. Happy to dig into any ones. You you you think would be it'll lead to an interesting direction. Sure. So I'll start with 2 and then good 1.

[Speaker: George Hotz]
[00:15:18] So 2, I moved to Berkeley in 20 14, and I threw myself at the Merry cult. I showed up at the Mary Office, and I'm like, hi. I'm here to join your cult. It was a 2 story. And what I started to realize was, Mary, And less wrong in general have a very poor grip on the practicalities of politics. Very much. I think there was sort of a split you know, Curtis Sharvin, like, meal reaction. This is a spin off of rationality. And it's a spin off of rationality that understood the truth about human nature. So when I give you that ex you give that example of 2 chimps meeting in the woods and they're gonna fight. If I'm 1 of those chimps, at least I stand a chance, right? He might beat my ass. I might beat his, but if I come up against the FBI, Yeah. Things do not look good for me. In fact, things so much do not look good for me. There's no way I'm gonna beat the FBI. The modern forces are so powerful that this is not a, oh, we've established a nice cooperative shelling point. This is a We have pounded so much beer into these people that they would never even think of throwing a punch or firing a gun. We have made everybody terrified And this isn't good. We didn't we didn't achieve this through some enlightened cooperation. We achieved this through a massive propaganda effort. Right? It's the joke about, you know, the American soldier goes over to Russia and it's like, man, you guys got some real propaganda here. And that the Russian soldiers like, yeah. No. I know it's bad, but it's not as bad as yours. And the American soldiers like, what

propaganda? And the Russian just laughs. Right? So so this this didn't occur because of this occurred because of a absolute tyrannical force decided to dominate everybody. Right? Now. Oh, I think so. I think there's a way out of this. I think there actually is a way out of this. Right? And I wrote a blog post about this called individual sovereignty. And I think a really nice world would be if all the stuff to live, food, water, health care, electricity, we're generateable off the grid in a way that you are individually sovereign. And this comes back to my point about offense and defense. Right? If I have a world where You don't want it to be extreme defense. You don't want every person to be able to completely insulate them. But you want like, okay, it takes a whole bunch of other people to gang up to take that guy out. Right? Like, that's that's a good that's a good balance. And the balance that we live in today is there is 1 pretty much a Unipolar world. I mean, thank god for China. But, you know, there's 1 there's 1 Unipolar world. You got the America and What are you gonna run? I'll pay taxes. I don't care if you live overseas. Right? So, yeah, my my point about the coordination is that if you're okay with solving coordination problems, by using a single a singleton superintelligent AI to to to make everybody cower in fear and tyrannize the future. Sure. You'll get coordination. Yeah. That works. That works. I'm the only guy with a gun, and I got 10 of you. I kinda named it all 10 of you, and you can all die. Or listen to me. Tries. So I'm I'm curious about, so I understand what you're saying. And I think you make some decent points, but, think I view the world a bit differently from you, and I'd like to, like, dig into that a little bit. So, like, who do you think is less afraid?

[Speaker: Connor Leahy]
[00:18:45] Someone living just a median person living in the United States of America or the median person living in somalia.

[Speaker: George Hotz]
[00:18:53] Sure. America, less afraid.

[Speaker: Connor Leahy]
[00:18:55] Well, that's kind of strange. Somalia doesn't have a government. They have much less tyranny. You're much more you can just buy a rocket launcher and just like live on a farm and just like, you know, kill your neighbors and no one's gonna stop you. So, like, how does that interact with your old people? Those who will trade liberty for safety deserve neither. That sorry. I don't understand. Could you elaborate a bit more?

[Speaker: George Hotz]
[00:19:17] In Somalia, you have a chance. In America, you do not. Right? I am okay. I would rather live in fear. I would rather be worried about someone shooting a rocket launcher at me than to have an absolutely tyrannical government. Just you know, just just like like like a managerial class. I'm not saying, by the way. I agree with you that these things are possible. I agree with you that the less wrong notion of politics is possible. I would love to live in these sort of worlds, but we don't. The the the the practical reality of politics is so much more brutal, and it just comes from a straight up instinct to dominate, not an instinct, you know, government by the people for the people. Branding.

[Speaker: Connor Leahy]
[00:20:02] Yeah. I mean, yeah. To to be clear, I very much do not agree with less wrongs views and politics and a bit of an outcast for how I view how conflict or EIVU politics. But this

is I I feel like you're kind of dodging the question here just a little bit. It's like, well, if that's true, why aren't you living in Somalia?

[Speaker: George Hotz]
[00:20:20] I know people who've done it. Right? It's very hard. It's very hard psychologically. Okay. So, like, tigers love chum. It turns Right? A tiger does not want to chase down an antelope. Right? A tiger would love to just sit in the zoo and eat the chum. Right? And, like, it takes a very strong tiger to reject And I'm not that strong. I hope there's people out there who are. I hope there's people out there who are actually, like, you know, I'm just not. I'm a weak little bitch. That's why it all comes to buy you. Right? Okay. I mean,

[Speaker: Connor Leahy]
[00:20:51] that's a fair answer, but I am a bit confused here. So you're saying that living in Somalia would be better by some metric, but you're also saying you prefer not living in Somalia. So I I am a bit confused because, like, from my perspective, I want to live in a country I want to live in, and that's the 1 which I think is better. If I thought that the other country was better,

[Speaker: George Hotz]
[00:21:13] then I would just move there. But person, let's the tiger and the chum, I think, is a good analogy. Right? Like, if you have a choice as a tiger, you can live in a zoo and you get a nice sized pen. You know, the zookeepers are not abusive at all. You get fed this beautiful chopped up food. It's super easy. You sit there and get fat, lays around all day, or you can go to the wild. And in the wild, you're gonna have to hunt. You might not succeed at hunting. It is just a, you know, it's a brutal existence. As a tiger, which 1 do you choose? Now, you say, oh, well, obviously, you know, you're going to choose the chum 1. Yeah, but do you see what you're giving up?

[Speaker: Connor Leahy]
[00:21:51] Do you see it? No. Could you elaborate a little bit on what I'm giving up? You are giving up on

[Speaker: George Hotz]
[00:21:57] the nature of Tiger You you are effectively. Okay. Maybe I'll I'll take this to an extreme, right? In the absolute extreme, The country that you would most rather live in is the 1 that basically wire heads you. Right? The 1 and you can say that. Okay. Well, I don't wanna be wireheaded, but, you know, there's a there's a gradient that'll get you there. Gandhiie in the pill. I you know, If you can live in this country, you can be happy, feel safe and secure all the time. Don't worry exactly about how we're doing it, you know. But, right? I mean, it takes a very strong person to it's going to take a very strong person to say no to wire head.

[Speaker: Connor Leahy]
[00:22:39] So I understand. Sorry.

[Speaker: George Hotz]
[00:22:41] I'll give I'll give 1 more instrumental reason for living in America versus living in Semalia. If I thought that America and Symalia were both like steady states, I might choose

somalia. I don't think that. I think that being here, I have a much better way of escaping this. Of escaping the constant tyranny that we're in, and I think a major way to do it is AI. Okay. I think that AI is is If I really if I had an AGI, if I had an AGI in my closet right now, I'll tell you what I'd do with it. I would have it build me a spaceship that could get me off of this planet and get out of here as close to the speed of light as I possibly could and put a big shield up behind me blocking all communication. That's what I would do if I had an AGI, and I think that's you know, the right move. And I have a lot better chance of building that spaceship right here than I do in somalia. Right? So I'll give an

[Speaker: Connor Leahy]
[00:23:29] that's alright. Man, that's a good instrument. Well, we'll miss you if you leave, though. That'll be real shame. It'll be

[Speaker: George Hotz]
[00:23:35] Everyone should do it. Like, this is this is the move. Right? And, like, let humanity blow I mean, look, I agree with you that we're gonna probably blow ourselves up. Right? But I think that the path potentially through this probably looks different from the path you're imagining. I think that the reasonable position I'm sorry. Oh, no. No. I think, yeah, maybe we're done with this point. I can come back and have a response to your first and last time. I would like to, if you don't mind, just, like,

[Speaker: Connor Leahy]
[00:24:02] full on 1 1 string there as well. So 1 of the things you said is, like, what will a tiger choose? And so my personal view of this kind of thing. And I think when I think about coordination is I think of things so you put a lot of view on this, like, fear based domination and so on. And I'm not gonna deny that this isn't a thing that happens. I'm German. You know? Like, you know, I have living relatives who could tell you some stories. Like, I understand. Like, I I understand. I'm not I'm not denying these things by me means. What I'm saying though is, okay, let's say it was a bunch of tigers. You know, you and me and all the other tigers, and some of the tigers are like, Man fuck. This whole, like, nature shit is like really not working for me. How about we go build a zoo together? Who's in? And then other people are like, yeah, you know what? Actually, that sounds awesome. Let's do that. Do you think that's okay? Like, you think that would be like a fair option for them to do?

[Speaker: George Hotz]
[00:25:00] Sure. But that's not where Zoos come from. I I know. I know. I'm I'm getting there. I'm getting there. Okay. So, like, that is not where Zoos come from. Sure. But

[Speaker: Connor Leahy]
[00:25:08] The the technology here is, of course, is that this is where a lot of human civilization not all of it. I understand that why France was doing well in the first World War was not because of democracy being nice. It was because democracy raises large armies. It's, I'm very well aware of realpolitik as the Germans would say -- Yeah. -- about these kinds of factors. And I, and I fully agree with you that a lot of the good things that we have are not by design, so to speak. You know, they're happy side effects, you know. Capitalism is a credit assignment mechanism. You know, the fact that also resulted in us having cool video games and air conditioning is not an inherent feature of the system. Them. It's it's it's an execution mechanism. And so totally grant all of this. I'm not saying that every coordination

thing is good. I'm not saying that, you know, there aren't trade offs. Especially, you were talking about, I think, aesthetic trade offs, you're like there's an aesthetic that the tiger loses by living in. Yes. Sure. And, well, I think it's personally aesthetics are subjective. So I think this is something that different So the way I think about aesthetics is I think aesthetics are things you trade on is, you know, you might want tigers in the wild to exist. Okay. Fair enough. That's a thing you can want. You know, someone else might want, you know, certain types of art to exist. They might want, a certain kind of religion to be practiced or whatever. These are aesthetic preferences upon reality, which I think are very fair. So the way I personally think about this morally, is I'm like, okay. Cool. How can we maximize trade surplus so you can spend your resources on the aesthetics you have you want and I'll spend my resources on the, you know, things I want. Now maybe the thing you describe where everyone just atomizes into their own systems, with their own value system, with their own aesthetic, completely separate father is the best outcome. Awesome. I think this is completely

[Speaker: George Hotz]
[00:27:01] Have you had the have you had the innovator manifesto?

[Speaker: Connor Leahy]
[00:27:04] I am not. You sure.

[Speaker: George Hotz]
[00:27:07] The problem with this, everyone trades on their own aesthetics is you will never be able to actually buy any aesthetics that are in conflict with the system. Right? The you won't? The system won't let you.

[Speaker: Connor Leahy]
[00:27:20] Okay. By by that logic, why do people have free time? Why don't they work all the time? Why doesn't capitalism extract literally every minute of them? Do you think that is?

[Speaker: George Hotz]
[00:27:33] I think it's because it turns out that we don't actually live in a capitalist society. I think China is a lot closer to a capital a society than America. I think America is kinda communist, and I think in a communist society, of course, you're gonna get free time. It turns out that subsidizing all the homeless people is a great idea. Right? If you wanna keep power, again, do some absolute tyrannical mechanism. You do it. Right? So why do we have free time Well, you think it's some victory of capitalism. I think it's because we do not live in a capitalist country. I think China's more capitalist than America. Think it's because we trade on our aesthetics. I think that different people have different things to contribute to various systems, and not necessarily capitalist or communist thing. I'm saying it's it's more energy.

[Speaker: Connor Leahy]
[00:28:13] Is that in the in the primordial environment, if you have to fight literally every single second and spend every jewel of energy you have to scrou scrounge together another jewel of energy, you can't have free time. It's not about capitalism. This is about an entropy. This is about these kind of things. We have energy access. We have we've produced systems that allow us to extract more energy per jewelry put in, and we can spend that extra energy on things such as free time. And the distribution of, you know, energy, power, coordination,

whatever you want to call it is, is another question. Will you agree or disagree with this? I mean, I am taking an extreme position when I say that There are definitely positive sum coordination

[Speaker: George Hotz]
[00:28:51] problems that are solved by governments. Right? It is not all 0 solve or negative sum. Right? I'm not I'm not denying this. But what I'm saying is it's like, I don't know, man. Like, the existence of free time Well, that's all great when you think you live in this surplus energy world. Right? And maybe we do right now. But if some other country took this seriously like China, Who's gonna win in a war? Who's gonna win? Is it gonna be the Chinese? You ever see the Chinese build a building? They got like 400 people there, and they're all there 24 hours a day, and they're getting the building built. You ever see Americans build a building? It's 6 guys. 2 of them are working. 2 of them are shift supervisors and 2 of them are on lunch breaks. Oh, you got your free time. You got your aesthetic preferences. You know, you deserve to lose in a war. Right? This country deserves to lose in a war if they keep acting the way they're acting.

[Speaker: Connor Leahy]
[00:29:42] So I I definitely see the point you're making. And this is personally not a thing I wanna defend too, fur, because I'm not a military expert, but I will note that I'm not a good expert either. I will note that the US has, like, 37 aircraft carriers and the Chinese have, like, 2, and Americans are, like, somehow, you know, despite being so lazy. And, oh, no, they have all this, you know, all this free time or whatever. Somehow, there's still military hegemon and whatever. And, like, the biggest rival Russia fighting this backwards water country in Ukraine suddenly fold and lose like 3 quarters of the military. It's What I'm saying is, if you have massive hegemony, if you have truly, obnoxious victory, the way it should look is that you lase around all the time and you look like a fucking idiot you still win?

[Speaker: George Hotz]
[00:30:26] Yes. And I'm not talking about Russia. Russia has a GDP the size of Italy. This is China here. You might say that China has 2 aircraft carriers in the US has 37. Why do we have aircraft carriers? Who has more drone building capacity? The Chinese are the United States. If the future is fought with AI swarm drone warfare, the Chinese can make, you know, a million drones a day, and the US can make I don't even know. I think we buy them from China.

[Speaker: Connor Leahy]
[00:30:54] I'm not an expert on these kind of logistics. I think I would like to get back to kind of like the more general Let's let's move on from that. I am not either, but I do believe the Chinese have more manufacturing capacity than the United States.

[Speaker: George Hotz]
[00:31:06] It seems completely plausible to me. I think things are complicated. I'm lazy and they don't sit around and have all this free time and aesthetic preferences or something. I'm a believer that work is life.

[Speaker: Connor Leahy]

[00:31:16] I mean, at least from my Chinese friends, I know, the Chinese sure do have a lot of inefficiencies. It's just called corruption.

[Speaker: George Hotz]
[00:31:24] Oh, America has corruption too. You see? Oh, yeah. Sure. Well, in Mexico, the corruption is you have to pay 20 cents to get, you know, 20 cents on every dollar for the building you Right? Whatever, man. In America, you've every dollar is spent absolutely on that building. You know how we know that because we spent 4 dollars making sure that that first dollar was not spent corruptly.

[Speaker: Connor Leahy]
[00:31:43] I I'm I'm I'm well aware of that, sir. Anyways, I would like to, like, I think I think we mostly agree on this point, actually, and I think it's a matter of degree. I I what I wanna say, just for the record, the US is a, like, uniquely dysfunctional system in the west. I'm German. And, like, the German system is very dysfunctional. But it's like nothing compared to how dysfunctional the US is. Fully agreed with that. I don't I don't think we disagree on that. I think it's a matter of degree more so than I agree. We've just had we've had a we've had a comment saying someone's turned the temperature up a bit too much on the language model. So let's bring it back a tiny bit to AI safety, but that that was a that was a great discussion. Got it. I will end up saying, I love America. I am happy to live here, and there are a lot of things I appreciate about American society. Great. So do you wanna return to, like, the technical topics? Or

[Speaker: George Hotz]
[00:32:31] Yeah. I think I can return to your first point, and maybe I'll just start with a question. Do you think there's gonna be a hard take off?

[Speaker: Connor Leahy]
[00:32:37] I don't know, but I can't rule it out.

[Speaker: George Hotz]
[00:32:42] I can't see how that would possibly happen.

[Speaker: Connor Leahy]
[00:32:46] I have a few ideas of how it could happen, but I don't it's, like, unlikely. It seems like not The the way I think it could happen is if there are just algorithms, which are like magnitudes of order better than anything be ever have. And, like, the actual amount of compute you need to get to human is, like, you know, a cell phone or, you know, like, and then this algorithm is not deep in the tech tree. We just happen to have not picked it up, and then an AGI system picks it up. This is how I think it could happen.

[Speaker: George Hotz]
[00:33:15] Okay. Yes. I I agree that something like this is potentially plausible. What you're saying, basically, like, the god shatters already distributed, the the, the, the, It's it's it's not a question. It's using all the existing compute in the world today. It just turns out it was 10000 x more effective or a million x more effective than we thought. Yeah. This just seems the most plausible way to be up. Or, you know, you mix lead and, you know, copper and you get a superconductor, you know, something like that. Even. I know. I know. I'm not joking. It's

gonna take so many years to like, it's not about the discovery. Right? Give it 10 years to productionize at scale up processes. Right? Like these things are you know, this is something running a company has really taught me. Like, it's just gonna take a long time. And this is really, like, like, kind of where my I just don't believe in a hard take off. I think that they'll be this is a Evgasket thing I like. He's, hardware and software progressed at quite similar speeds, and you can look at factoring algorithms to show this. So it would shock me if there were some, you know, 10 to the 6, 10 to the ninth magical improvement to be had.

[Speaker: Connor Leahy]
[00:34:17] It seems plausible to me. Like, a hard take off is definitely not my main line scenario. My main line scenario well, I don't know. Maybe you wouldn't consider this a heart maybe you would consider this a heart attack off. This is what I would describe soft take off is something like sometimes the way I like to define AGI is say it's, something that has the thing that chimp that chimbs don't have and humans do have. Yeah. So chimbs don't go a third to the moon. You know, despite their brain being a third of our size. So we scaled up things by a factor of 3 of a primate brain roughly or 4 or something like that. And, like, most of the structure is same. Sure. Some micro tweaks and whatever, but, like, not massive amount of evolutionary pressure. Like, we're very, very similar to chips. And somehow, this got us from you know, literally no technology to space travel in a, you know, evolutionary, very small period of time. It seems imaginable to me that something similar could happen with AI. I'm not saying it will, but, like, seems to match

[Speaker: George Hotz]
[00:35:17] Yeah. So I agree with this. I'll I'll come to your point about, you know, you had 2 regulatory points. 1 of them about, capping the max flops. And I actually kind of agree with this. I do think that things could potentially become very dangerous at some point. I think your numbers are way way way too low. I think if your numbers are anywhere near GPT 3, GPT 4, Okay. Great. We got a lot of we got a lot of fast moving guys who work on fiverr, even if you start to get von Neumann's. Right? We're not talking about a humanity's worth of compute. We're talking about things on par with a human and a few humans. Right? You have to run fast, but they're not Like, things get scary when you can do a humanity's training run-in 24 hours. Like, we're about to burn the same compute that that all 2000000 years of human civilization burned. Okay. Now I don't know what starts to happen, or I'll put this kind of another way. Language models, I look at them, and they don't scare me at all because they're trained on human training data. Right? These things are not Like, if something was a good as as good as GPT 4 that looked like new 0, where it trained from some simple rules, Okay. Now I'm a bit more scared. But when you say, okay. It's, you know, or we're feeding the whole internet into the thing and it parrots the internet back mushed around a little bit. That looks very much like what a human and I'm just not scared of that.

[Speaker: Connor Leahy]
[00:36:44] Yeah. It's very reasonable. Whatever. Like, I, I, I, I'm not scared of Jeep before to be clear. Like, I, I think there is, like, 0 percent chance or, like, epsilon chance that g p t 4 is existentially dangerous by itself. You know, maybe some crazy g p t 4 plus r l plus 0 plus something something maybe. But I definitely agree with you here. I don't expect you to be 3 or 4 by themselves to be dangerous. These are not I'm much closer to, I think, what you were saying, like, yeah. If you had a mu 0 system, that boots drafted off GPU for

[Speaker: George Hotz]
[00:37:13] holy shit. Like, we're a big, we're a big big shit if we get to it. Well, then then we should let's stop. Let's stop. Yeah. Let's let's let's stop. So so

[Speaker: Connor Leahy]
[00:37:20] I'm very happy to get it to be into a regime where we're like, okay, let's find the right bound. Like, I think this is an actually good argument. I think this is actually something that should be discussed, which is not obvious. And I could be super wrong about that. So I'd like to justify a little bit about why I put such an small bound, but I think the arguments you're making for the higher bounds are very reasonable I think these are actually good arguments. So just to justify a little bit about why I put such a low bound, the boring default answer is conservatism. Is like if all of humanity is at stake, which, you know, you may not believe. I'm like, whoa, whoa. Okay. At least give us a few years to, like, more understand what we're dealing with here. Like, I understand that, you know, you may disagree with this. Very plausible. But I'm like, Well, like, you know, at least let's let's, like, by default, let's hit a pause button for, like, you know, couple of years until we figure things out more. And then if we, like, find a better theory of scaling. We understand how intelligence scales. We understand how mu 0 comes, blah, blah, blah, and then we pick back up after we're like, you know, or we make huge breakthroughs in alignment. And eliezer is is crying on CNN and like, oh, we did it, boys. I mean, then, okay. Sure. You know, okay. So that's the 1, like, kind of more boring argument. Like, that's kind of the boring argument. The more interesting argument, I think, which I I think is a bit, you know, more schizo. Is that it's not clear to me that you can't get dangerous levels of intelligence with the amount of compute we have now. And 1 of the reasons that I'm I'm unsure about this is is because, man, GB 3, GPforce is the dumbest possible way to build AI. Like, it's just, like, like, there's like no dumber way to do it. Like, it's it works and dumb is good, right, you know, bitter or less than dumb is good. But Look at humans. You use, as we talked about before, you know, human today, human 10000 years ago, not that different. You place both of them into a, you know, workshop with tools to build, you know, any weapon of their choice, which of them is more dangerous. Obviously, you know, 1 of them will have much better, you know, capacities

[Speaker: George Hotz]
[00:39:25] to deal with tools, to read books, to

[Speaker: Connor Leahy]
[00:39:28] think about how to design new weaponry and so on. These are not genetic changes. They are epistemological changes. They are memetic. Their software updates. You know, humans had to discover rational reasoning. Like, you know, before, like, you know, I mean, you know, obviously, people always had, like, folk conceptions of rationality, but it wasn't like a common thing to think about causality and like, you know, you know, rational, like, you know, if then else kind of stuff until relative, you know, like philosophers in the old ages and only became widespread relatively recently. And these are useful capabilities that turned out to be very powerful and took humans many, many thousands of years to develop and distribute. That's good. And I don't think humans are anywhere near the level. I think the way we could do science right now is pretty awful. Like, it's like the dumbest way to do science that, like, kinda still works. Like, you know, and I expect it's, like, possible that if you had a system, which, like, let's say it's, like, smaller brain than a human even, but it has really, really sophisticated histomology. It has really, really sophisticated theories of meta

science, and it never tires. It never gets bored. It never gets upset. It never gets distracted. And it can, like, memorize arbitrary amounts of data. This is something that I think is within the realm of like a GPT 3 or 4 training run to build something like this. And it is not obvious to me that this system could not out point humanity. Maybe not. Like, maybe not. But it's not obvious to me that it can't. So,

[Speaker: George Hotz]
[00:40:59] just curious what do you think of that? So to your first point, why I stand against almost all conservative arguments. You're assuming the baseline is no risk. Right? And, oh, well, why should we do this AI? We should wait bring the baseline back. No. No. No. No. No. We are about to blow the world up any minute. There is enough nuclear weapons aimed at everything. This is wearing some incredibly unstable precarious position right now. Like, people talk about this with with car accidents, you know, as comma, like, People are like, oh, well, you know, if your device causes even 1 accident, I'm like, yeah, but what if statistically there would have been 5 without the device? I'm like, You do have to understand the baseline risk in cars is super high. You're making 5 x safer. There's 1 accident. You don't like that? Okay. Mean, you have to be excluded from any polite conversation. Right? Right? So, yeah, like, I I think that calling for a pause to the technology is is, worse. Right? I think given the 2 options, if we should pause or we should not pause, I think pausing actually prevents more risk. And I can talk about some reasons why. Again, the things that I'm worried about are not quite The existential risks I have to the species are not AGI goes rogue. They are government gets control of AGI and ends up in some really bad place where nobody can compete with them. I don't think these things look unhuman. These things to me, like, I see very little distinction between human intelligence and machine intelligence It's all just on a spectrum. And, like, they're they're not, like, to come to the point about Okay. But GPT 4 could be like this hyper rational, never tiring. Humans are doing science in the dumbest way. I'm not sure about that. Right? Like, I think that, you know, when you look at, like, okay. Well, okay. We have chess bots that do way better. And all they do is think about chess. Haven't really done this with humans, people would call it unethical. Right? Like, if we really told a kid, like, if we really just, like, every night. We're just putting the chess goggles on you and you're staring at chess boards and we're really just training your neural net to play chess. I think humans could actually be the computer again at chess if we were willing to do that. So yeah, I don't think that this stuff is that particularly dumb. And I think, okay, maybe we're losing 10 X, but we're not losing a million X. Again, I don't see a I do the numbers out all the time for when we're gonna start to get more computer, you know, when will a computer have more compute than a human, when will a computer have more compute than humanity. And, yes, these things get scary, but we're nowhere near scary yet. We're looking at these cute little things. And These things, by the way, do present huge dangers to society. Right? The psi ops that are coming. Right now, you assume that, like, when you call somebody, that you're at least wasting their time too. But we're gonna get like, heaven banning. I love this concept, which is, you know, yeah. Yeah. Okay. Up on Luther ai. Like, that's where it comes from. I was like, yeah, on a Luther ai that came up with that word.

[Speaker: Connor Leahy]
[00:44:10] You guys did? Oh, yeah. Yeah. I know the guy came up with it. I I I love I love this concept, and I think, there's also a story, to to, my little pony friendships optimal.

[Speaker: George Hotz]

[00:44:20] That that god that goes into the concept. And yeah. So I think that, like, my, my girlfriend proposed a, I don't wanna talk to oh, say you don't wanna talk to your relative anymore. Right? Mhmm.

[Speaker: Connor Leahy]
[00:44:37] Okay.

[Speaker: George Hotz]
[00:44:38] Give my AI version to talk to. Right?

[Speaker: Connor Leahy]
[00:44:41] Yeah.

[Speaker: George Hotz]
[00:44:42] Rest. Yeah. So, like, this stuff is coming and it's coming soon. And if you try to centralize this, if you try to, you know, say like, oh, well, okay. Google OpenAI, great. They're not aligned with you. They're really not. Google has proven time and time again. They're not aligned with MedA has proven time back on time again. They're trying to fix it, but, you know.

[Speaker: Connor Leahy]
[00:45:01] Yep. I mean, I I fully agree with you. Like, I like that you bring up psi ops as the correct example in my opinion of short term risks. I think you're, like, fully correct about this. Like, when I first saw, like, GPT models. I was like, holy shit, like, the level of control I can gain over social reality using these tools at scale is insane. And I'm surprised that we haven't seen yet the things that, like, augured in my visions of the day, and we will. Like, we will, obviously, it's coming. And This is so I think this is a very, very real problem. Yeah. Like, I think if we even if we stop now, we're not out of the forest. So, like, so, when when you say, like, I, I think the risk is 0. Please do not believe that that is what I believe because it is truly not. It is truly truly not. I think we are like, we are really in a bad situation. We are in a we are being we are under attack from, like, so many angles right now. And this is before we get into, you know, like, you know, potential, like, you know, climate risks, nuclear risk, whatever, we're in under rheumatic risk. Like, the the dangers of our, like, epistemic foundations, are under attack. And this is something we can adapt to. Right? Like, you know, we did, you know, when, a good friend of mine, he is, he's quite well read on, like, Chinese history. And he always like it tells me his great stories. So I'm not his story, and so please, you know, don't crucify me here. But, like, tells these great stories about when Marxist memes were first introduced to China. And, like, this is where a world where, like, just, like, all the precursor news didn't exist. That's just like kind of was air dropped in. And people went nuts. People went just completely crazy because there was no memetic antibodies to these like hyper virulent means that were, you know, created by evolutionary pressures in, like, you know, western university departments. Like, really, you could call a philosophy department just gain a function in the medical laboratories.

[Speaker: George Hotz]
[00:46:49] I like that. I like that. Yeah. That's that's what you are.

[Speaker: Connor Leahy]

[00:46:53] I mean, like, you know, like, without being, you know, political or any means there, a lot of what these organizations do. And, like, you know, other, you know, what, other, you know, memetic, like, you know, if philosophy departments are the, like, gain of function laboratories, then, like, fortune and Tumblr are, like, the bat caves of meats, you know, like the Chinese bat caves. And I I remember this vividly. I was, like, on on Tumblr and 4 and 4 chan, like, when I was a teenager. And then suddenly all the, like, weird bizarre, you know, internet shit I saw started becoming mainstream news might parents were watching in 2016. And I was like, what the hell is going on? Like, I already developed antibodies to this shit. Like, I already, you know, both right and left. That was already, like, I already immunized all of this. So I fully agree with you that this is, like, 1 of the largest risks that we are faced is this kind of like mimetic mutation load in a sense? And I'm not gonna say I have a solution to this problem. I'm like, I have ideas, like, there's a lot of, like, things you can do to improve upon this. Like, if AI was not a risk and also not climate change and whatever, this might be something I work on, like epistemic security. This might be something I would work on. Like, how can we build better coordination like like just scalable rationality mechanisms, stuff like prediction markets and stuff like this? I don't know. But sorry, going off track here a little bit, but Well, no. Actually, I I really

[Speaker: George Hotz]
[00:48:14] agree with a lot of the stuff you said, and I had similar experience with the antibodies and people are exposed to this stuff. And I'm like, yeah. This got me, like, 4 years ago. Yeah. So I think that there is a solution, and I have a solution, and the answer is open source AI. The answer is open source Let's even you can even dial it back from, like, the political and the terrible and just straight up talk about ads and spam. Or maybe spam, just straight up spam. I get so much spam right now. And it's like, it's kinda written by a person. It's like targeting me to do something, and Google's spam filter can't even come close to recognizing it. Right? Like, what I need is a smart AI that's watching out for me that is just it's not even targeted attacks at me. It's just so much noise. And I don't see a way to prevent this. Like, the big organizations, they're just gonna feed you their noise. Right? And they're gonna maximally feed you their noise. The only way is if you have an AI, like, I don't think alignment is a hard problem. I think if you own the computer and you run the software, if you develop the software, the AI is aligned with you. Oh, yeah.

[Speaker: Connor Leahy]
[00:49:21] Can you okay. If I -- Yeah. -- challenge you, George Hoss, here is a llama 65 b model when we computer to run on. I didn't know that. Make it so it make, yeah, you know, sure. You okay. You developed it. I I give you the funding, your time. Can you develop a model that is as good as Islam c 4 B, 6 5 B, and it's immune, like, completely immune to jail breaks. It cannot be jailbroken. No. Why not? It's aligned, isn't it?

[Speaker: George Hotz]
[00:49:46] Well, no. But this isn't what alignment means. Well, my values is do not get jailbroken. Oh, okay. You're talking about unexploitability. This is not alignment. Right? Oh, okay. Okay. Interesting. I didn't know you would separate those. Oh, right. Extremely separate those. Right? Okay. Interesting. It means in the default case, it like, like, it it's on my side. Right? Mhmm. Unexploitability is not a question of whether it's okay. This is a true thing about people too. Whenever I look at a person, I ask, okay, is this person I want something from you? Is this person does this person want it to? And is this person capable of doing it?

Right? And I really separate those 2 things. I can build you a system. I don't I'm not worried about the first 1 with the AI system. I'm worried about the second 1. Can it be gamed? Can it be exploited? Sure. I could tell, like, you know, like, like, say it was just playing chess, right? And it loses. I'm like, don't lose Okay? I didn't want to man. I didn't wanna lose. I'm sorry. I know. But, like, so yes. Yes. Can I build a aligned system? Sure. Can I build an unexploitable system? No. Especially not by a more powerful intelligence.

[Speaker: Connor Leahy]
[00:50:52] Interesting. Interesting. This is an interesting I I think you're you're you're appointing to actually a very important part of this. Is that like exploitability and alignment can get fuzzy, like, which is which? Like, did it fail because of its skill set or because it's not aligned? It's actually a very deep question. So I think I think you make a good point for, like, talking about these 2 separately. I guess, the, so the thing I want to dig in just like a a little bit more on on this idea is there are there's 2 ways. There are 2 portals through which, you know, the memetic demons can reach into reality, humans and computers. Why do you think your AI is immune to means? Why why can't I just build AIs that target your AIs? Well, like you Don't. I don't think my AI is immune to memes at all.

[Speaker: George Hotz]
[00:51:40] I think that the only question is and I really like your key. Like, these these NGOs are doing gain of function on memes. Right? Where are masks? The like a a weaker intelligence will never be able to stand up to a stronger intelligence. So from this perspective, if this is what's known as alignment, I just don't believe that this is possible. Right? Because you can't you can't keep a stronger intelligence in the box. This is this is I agree with you cowskin in the box experiments. It's like the AI is always going to get out. There's no keeping it in the box. Right? This is this is a complete impossibility. I think there's only 2 real ways to go forward. And 1 is Tech Kaczynski. 1 is technology is bad. Oh my god. Blow it all up, let's go live in the woods. Right? And I think this is a philosophically okay position. I think the other philosoph the okay position is something more like effective accelerationism, which is, look. These AIs are going to be super powerful. Now, if you have 1, it could be bad. But if superintelligent AIs are all competing against each other, memetic Like, we have something like society today, just the general power levels have gone up. This is fine as long as these things are sufficiently distributed. Right? Like, Sure. This AI is not perfectly aligned, but, you know, there's a thousand other ones, and, like, you have to assume they're all basically good because they're all basically bad while we're dead anyway. I mean, why wouldn't you expect that? That they're all bad? Yeah.

[Speaker: Connor Leahy]
[00:53:13] Well, or what do you think of humans? Are most humans good? Yeah. Most humans? I think the concept of good that doesn't really apply to humans because humans are too inconsistent to be good. Like, by default, they they can be good in various scenarios in various social contexts. Like, give me any human, and I can put them into a context where they will do an arbitrarily bad thing.

[Speaker: George Hotz]
[00:53:31] And this is true about Alama as well. Right? Lamas are completely inconsistent. I think they're actually more inconsistent than humans. Right? Yeah. But I wouldn't trust Lamas to be good. Well Yeah. But I wouldn't think that they're bad either. I would think they

have the exact same inconsistency problem as humans, and I think almost any any AIU build is gonna run into these same problems. Right? Yeah. No 2 reason to think. So that's that's my point. So your your assumption can't rely on them being good because you don't get that for free. Like, where does that come from? My assumption is not that they're good, my assumption is that they're not bad. But inconsistent is fine. As long as we have ton of them, and they're all inconsistent, and they're pulling society in every which direction. You don't end up paper clipped, right?

[Speaker: Connor Leahy]
[00:54:11] Why not?

[Speaker: George Hotz]
[00:54:12] Well, because What they're all gonna coordinate and agree to paper clip you? No. No. They'll just do some random bullshit, and then that random bullshit will not include humans. They're all doing random bullshit. Right? You're gonna have let's say the liberals decide we're gonna paperclip people. The conservatives are gonna come out very strongly against paper clipping. Right? Like, and you're just you're just gonna end up with these sort of dynamics, like, forever, or you're gonna have some ais who are like, yeah, we don't like it, you know.

[Speaker: Connor Leahy]
[00:54:40] So this is perplexing to me because, like, in my view, which I think I heard you agree with earlier, it's like the world is unstable. The world is very unstable. And it seems to me your suggestion if we increase the chaos, if we have more entities fighting, doing more different things, with more energy that makes it more stable. Is that correct?

[Speaker: George Hotz]
[00:55:00] The world has actually become very stable in the last hundred years, and I'm scared of how stable it's become. It would, you know, again, thank you for China. How many sovereign countries are there in the world?

[Speaker: Connor Leahy]
[00:55:12] I mean, you know, hundred 90

[Speaker: George Hotz]
[00:55:14] something that's the UN number if you believe them. I think there's the US and China and maybe Russia. Mostly, they just have a veto button. Right? So I think there's more like 2 and a half. Like, this isn't a lot of entities fighting it out, chaionic. A lot of entities fighting it out chaotically would be intelligence is spread across the entire galaxy. Right? And that's the other that's the other beautiful thing too. Sooner we get off this planet, the sooner we get things that are actually a light year away, well, I think the speed of light's real. It's gonna take you a year to send them their your your scary memes. Right?

[Speaker: Connor Leahy]
[00:55:50] You want to get away from the memes? There's only 1 way. That's, you know For for the record, I would like to go on the record and if any future transhumanists or whatever want to blast himself into space and go do their own thing, I support their their right to do that. And I would love to give this right to people. The number 1 thing I want from countries is

the ability, the the right to leave. This is what I would love. This is what I love about companies. Free, I guess you have it. You're talking. You're talking. New reaction Tom. Yep. Yeah. Free exit is extremely important. I would not describe myself in the reactionary, please, because I'm not that gay, but,

[Speaker: George Hotz]
[00:56:23] and and I wouldn't describe myself that way either, but I've heard a lot of good ideas from them.

[Speaker: Connor Leahy]
[00:56:29] But, yeah, that being said, I do I do think that, like, you know, what I want like, I think let let's ground the conversation, like, a little bit here. So, like, I think I I love, like, the I'm very enjoying this conversation. I love talking these philosophical points. I think these are really good points. Really interesting. But ultimately, you know, as we all it to, like, the, you know, ladder third of this conversation. The thing I really care about is strategy. Okay. The thing I really care about is we are politics. I really care about Okay. What action can I take to get to the features I like? Yep. And, you know, I I'm not, you know, gonna be 1 of those galaxy brain fucking utilitarian, like, well, actually, this is the I'm like, no. No. This is what I want. Look, I I like my family. I like humans. You know, look, I, yeah, it's just what it is. Right? Like, I'm not gonna justify this on some global beauty, whatever. It doesn't matter. So I wanna live in a world. I wanna I wanna in 20 years time, 50 years time. I wanna be in a world where, you know, my friends aren't dead. I'm like, where I'm not dead. You know, maybe we are like, you know, cyborgs or something, but I don't wanna be dead. So what I really care about ultimately is how do I get those wrong? And I want us all to not be offering. Right? Like, you know, I don't wanna be in war. I want us to be, like, in a good outcome. So I think we agree that we would both like a world like this. And we think we probably disagree about how best to get there. And I'd like to talk a little bit about, like, what can we what should we do and, like, why do we disagree about what we Does that sound good to you? Maybe I'll first propose a world that meets your requirements.

[Speaker: George Hotz]
[00:57:57] And you can tell me if you want to live in it. So here's a world. We've just implanted electrodes in everyone's brain and maximize their reward function. I would hate living in the world like that. Yeah. But no 1 it meets your requirements, right? Your friends are not dead. No one's suffering and we're not at war.

[Speaker: Connor Leahy]
[00:58:14] That is true. There are more criteria than just that. I but the true the criteria I said is things I like. As I said, I'm not a utilitarian. I don't particularly care about minimizing suffering or maximizing What I care about is this various vague aesthetic preferences over reality. I'm not pretending this is I thought that was the whole spiel I was trying to make. Is that I'm not saying I have a true global function to maximize. I say I have various aesthetics.

[Speaker: George Hotz]
[00:58:40] I have various meta preferences of those not asking for a global 1. I'm asking for a personal 1. I'm asking for a personal 1. I'm asking for a personal 1 that you I don't care about the rest of the world. I gave you mine. I gave you what I would do if I had an NGI. Yep. So I'm getting on this rock, speed of life, as fast as I can.

[Speaker: Connor Leahy]
[00:58:56] Fair enough. I think if that is I would like to live in a world where you could do that. This would be a a feature of my world. If you a world where I would be happy is a world in which we coordinated around, you know, at at larger scales around building aligned AGI that could then distribute, you know, intelligence and matter and energy in a, you know, well hap value handshaked way between various people who may want to coordinate with each other may not. Now, some people might want to form groups that have shared values and share resources, others may not. I would like to live in a world where that is possible. Have you read Better Mortgage? I have unfortunately not.

[Speaker: George Hotz]
[00:59:39] Yeah, I was gonna ask you if you're happy with that world. Right? Like, like, unfortunately, don't know it. I I mean, Yeah. It's simple to describe. Singleton AI that basically gives humans whatever they want like, maximally libertarian, you know, you can do anything you want besides harm others. Is that a good world?

[Speaker: Connor Leahy]
[01:00:00] Probably. I don't know. I I haven't read the book. I assume the book has some dark twist about why this is actually a bad world. Not really. Not really. I mean, the the plot is pretty obvious. You are the tiger eating chong. Right?

[Speaker: George Hotz]
[01:00:12] Sure. But you can then just decide if that is what you want, then you can just return to the wilderness. That's the whole point. Yeah. But can you? Can you really return to the wilderness? Right? Like, like, like, like, you think that, like, I don't think we have free will. I don't think you ever will return to the wilderness. I think a large majority of humanity is going to end up wire head. Yeah. I expect that too. K. Great. And this is the best possible outcome, by the This is giving humans exactly what they want.

[Speaker: Connor Leahy]
[01:00:38] Yep. Yeah. Well, to be to be clear, I don't expect it's all humans. I truly do I think a lot of humans have meta preferences over reality. They have preferences that are not their own sensory experiences that thing a thing that the utilitarians get very wrong is that many human preferences are not about their own not even though they're not even about their own sensory input. They're not even about the universe. They're about the trajectory of the universe. They're about 4 d 4 d utilitarianism, you know? And a lot of people want struggle to exist, for example. They want heroism to exist or whatever. I would like those values to be satisfied to the largest degree possible, of course. Am I gonna say I know how to do that? No. Which is why I I kind of like didn't want to go this deep because I think if we're arguing about, oh, do we give them, you know, for the utilitarianism versus libertarian,

[Speaker: George Hotz]
[01:01:34] Utah, whatever. I mean, we're already, like, 10000 steps to I'm asking about you. I'm not asking about them. I'm asking about a world you want to live in. And this is a really hard problem. Right? Yeah. And this is why I just fundamentally do not believe in the existence of AI alignment at all. There is no there is no like like, what values are we aligning it to? Whatever the human says or what they mean or, like, Sure. Sure. But, like, my point is

I feel we have wandered into the philosophy department instead of the politics department. Okay. Like, It's like, I agree with you. Like, do human values exist? What does exist to me? But, like, by the point you get to the point where you're asking, what does exist to me, you've gone too And, like, I'll respond concretely to the 2 political proposals I heard you stayed on bankless. Sure. Okay. I'd love to talk about them. 1, is limiting the total number of flops.

[Speaker: Connor Leahy]
[01:02:30] Temporarily.

[Speaker: George Hotz]
[01:02:31] Temporarily. Yes. And what I I have a proposal for that, but I don't wanna set a number. I wanna set it as a percent. I do not want anybody to be able to do a 51 percent attack on compute. If 1 organization acquires 50 it's straight up 51 percent attack If 1 organization acquires 51 percent of the, compute in the world, this is a problem. Maybe we'll even cap it at something like 20. Know, in Canada, more than 20. Right? Yeah, I would support regulation like this. I would I don't think that this would cripple a country. But we do not want 1 entity or especially 1 training run to start using a large percentage of the world's compute, not a total number of flops. I mean, absolutely not. Cool. That'll be terrible. Like, we can actually agree. I would actually support that regulation. Like, nope, sorry, Sam Walton. You cannot 51 percent attack the world's compute. Sorry to leave them.

[Speaker: Connor Leahy]
[01:03:22] That's fair enough. I think this is a sensible way to think about things, assuming that, software is fungible, is that everyone has access to the same kind of software and that you have an offense defense balance. So in my personal model of this, I think, well, a, some actors have very strong advantages on software, which can be very, very large as someone who has trained very, very large models and knows a lot of the secret trick that goes into them, a lot of the stuff in the open source is far behind. Maybe we should force it to be open source. Well, this is your this is actually a very legitimate consequence for what I just said. And now I will say the second point about why I think that doesn't work. So the next reason why I think that doesn't work is that there is a there are constant factors at play here, is that the world is unstable. We've already talked about this. I think the amount of compute you need to break the world currently is below the amount of compute that more than a hundred actors actors have access to if they have the right software. And if you give if you have, let's say you have this insight, right, that could be used, nothing it will be, but it could be used to break the world, to, like, cause World War 3 or, you know, or just like, you know, cause mass extinction or whatever if it's misused. Right? Let's say you give this to you and me,

[Speaker: George Hotz]
[01:04:42] do you expect we're gonna kill everybody? Like, would you do that? Or would you be like, hey, let's, Hey, Connor. Let's, like, not kill the world right now, and I'll be like, Let's not kill them. How are we killing the world? How did we go from I I don't even understand, like, how exactly does the world get killed? This this is a big leap for me. I agree with you. I agree with you about the CIO stuff. I agree with you about Sorry. Sorry. Let let me you're right. I made too big of a over there. You're completely correct. Sorry about that. So to

[Speaker: Connor Leahy]

[01:05:10] back up a little bit. Let's assume we you and me have access to something that can train, you know, at mu 0 you know, super GPT 7 system on a tiny box. You know? Cool. Great. Problem is we do attach running with it, and we have it immediately starts breaking out and we can't control it at all. Breaking out. What was it? Yeah. I don't it immediately tries to maximize. It it learns some weird during the training process that is trying to maximize. And for some reason, this proxy involves gaining involves gaining, you know, mutual information about few about future state.

[Speaker: George Hotz]
[01:05:44] How is it gaining power? There's lots of other powerful AIs in the world who are telling it no. Well, we're assuming in this case it's only human Wait. This is a problem. No. No. No. No. You've you've ruined my entire assumption. As soon as it's you and me, yes, we have a real problem.

[Speaker: Connor Leahy]
[01:06:00] Chicken man is only a problem because there's 1 chicken man. Yeah. Yeah. I I look, I am with you. So I'm seeing before we get to the distributed case. So this is the the step before. We we it has not yet been distributed. Just, you know, you and me discover this algorithm in our basements. Okay. And so we're the first 1 who had it just by definition because know, we are the 1 who found it. What now? Like, do you think posting what do you think happens if you post this to GitHub?

[Speaker: George Hotz]
[01:06:25] Well, good things for the most part. Interesting. I'd love to hear more. Okay. So first off, I just don't really believe in the existence of we found an algorithm that gives you a million x advantage. I believe that we could find an algorithm that gives you a 10 x advantage. But what's cool about 10 x is like it's not gonna massively shift the balance of power. Right? Like, I want power to stay in balance. Right? This is, like, avatar, the last air pen. Power must stay in balance. The fire nation can't take over the other nations. Right? As long as power relatively stays in balance, I'm not concerned with the amount of power in the world. Alright. Let's just get to some very scary things. So What I think you do is yes. I think the minute you discover an algorithm like this, you post at the GitHub because you know what's gonna happen if you don't? The feds are gonna come to your door. They're gonna, take it. The worst people will get their hands on it if you try to keep it secret. So,

[Speaker: Connor Leahy]
[01:07:20] okay. That's a fair question, though. So I'll I'll I'll I'll take that aside. So am I correct in thinking that you think the feds are worse than serial killers in prison?

[Speaker: George Hotz]
[01:07:31] No, but I think that yeah. Well, yes and no. Do I think that your average Fed is worse than your average serial killer? No. Do I think that the feds have killed a lot more people than serial killers? All combined? Yeah. Sure. Totally agree with that. Not not not not It's not 1 fed. It's all the beds in their little in their little super powerful system. Sure. That's completely fine by me. Happy to grant that. Okay. What I wanna work run through is a scenario. Okay. Let's say, okay. You know,

[Speaker: Connor Leahy]

[01:07:59] we have a 10 x system or whatever, but we hit the chimp level. You know, we we we we jump across the chimp general level, or whatever. Right? You have a system, which is like John von Neumann level, whatever. Right? And it runs in 1 tiny box. And you get a thousand of those. So it's very easy to scale up to a thousand x. So, you know, So then, you know, maybe you have your thousand John von Neumann, improve the efficiency by another, you know, 2 5 10 x. You know, now we're already at 10000 x or a hundred thousand x, you know, improve Right? So, like, just from scaling up the amount of hardware, including with them. So

[Speaker: George Hotz]
[01:08:31] just saying, okay.

[Speaker: Connor Leahy]
[01:08:33] Now, Feds bust down our doors. Shit. You know, real bad. They take all our honey boxes. They're taking all of on newmans. They're taking all of on newmans. We're in deep shit now. We're getting chicken boys Shits are getting chicken. So okay. We get chicken, right? Bad scenario. Totally agree with you here. This is a shit scenario. Now the feds have you know, all of our AI's bad scenario. Okay. I totally see how this world goes to shit. Totally agree with you there. You can replace the Feds with Hitler. It's interchanged. Sure. But, like, I wanna, like, ask you a specific question here. And this might be, you know, you might say, nah, this is, like, too specific to each 1, but I wanna ask you a specific question. Do you expect this world to die is more likely to die

[Speaker: George Hotz]
[01:09:15] or the world in which the, you

[Speaker: Connor Leahy]
[01:09:18] know, Yak death cultists on Twitter who literally want to kill humanity who say this. Like, not all of them, there's a small subset of them. Small subset of them who literally say, oh, you know, the glorious future AI race should replace all humans. They break in, you know, with, like, you know, katanas and, you know, steal area. Which 1 of these you think is more likely to kill us?

[Speaker: George Hotz]
[01:09:43] Genuine question. To kill all of us, the feds, to kill a large majority of us, the EAC people.

[Speaker: Connor Leahy]
[01:09:50] Interesting. I would be really interested in hearing why you think that.

[Speaker: George Hotz]
[01:09:55] Sure. Oh, okay. So Actually, killing all of humanity is really, really hard. And I think you brought this up before. Right? You talked about, like, if you're gonna end up in a world of suffering? A world of suffering requires malicious agents where a world of, death requires maybe an act right? But I think this is plausible, but I actually think that killing all of humans, at least for the foreseeable future, is going to require malicious action too. Right? And I also think that, like, the fates that look kind of worse than death, like, I think mass wire heading is a fate worse than big war and everyone dies. Right? Like, like, a mass wire heading, like, a,

like, a singleton, like, a paper clipping, like, And I think that that is the 1 that the 1 world government and, you know, NGL New World Order people are much more likely to bring about than Yac. EAC, you're gonna have a whole lot of YAC people. Again, I'm not YAC. I don't live that my Twitter, but I think a lot of those people would be like, Yes. Spacehips. Let's get out of here. Right? Versus the feds are like yeah. Spacehips. Yeah. I don't know.

[Speaker: Connor Leahy]
[01:11:10] Interesting. So I think this is a fair opinion to the world. And It'll be outside our jurisdiction. How will we get taxes? I'm I'm describing more a very small minority of Yac people who are the ones who specifically goal their their anti natalist misanthropes. They want to kill the humans that is through stated goal. That they want humans to start, like, or, like, take extreme vegans if you want, you know, like, the, like, you know, like, my my argument, my point here I'm making is I'm not making the point feds are good by any means. I'm not saying it. What I'm saying is is that I would actually be somewhat surprised to find that the feds are anti naturalists who want to maximize the death of humanity. Like, maybe you have a different view here, but I find that knowing many Feds, that's quite surprising. I don't think that's what Feds want. Yeah. It's okay. So cool. So what you've so you you do agree that if we would post as open source, more of the insane death cultists, would get access to potentially lethal technology?

[Speaker: George Hotz]
[01:12:11] Well, sure. But again, like, it's not just the insane death cultists. It's everybody. And we as a society have kind of accepted.

[Speaker: Connor Leahy]
[01:12:20] It turns out everybody gets access to signal. Some people who use it are terrorists. I think signal is a huge good in the world. I agree. I fully agree with that. So, okay. Cool. So we've granted this that, you know, if we distributed widely, it would be given to some, like, incorrigibly

[Speaker: George Hotz]
[01:12:37] deadly lethal people. They're coordinating bombings on signal right now.

[Speaker: Connor Leahy]
[01:12:42] Sure. Sure. And then so now this this reduces the question. To a question about offense defense balance. So in a hypothetical world, which I'm not saying is the world we live in, but like let's say the world would be often favored such that, you know, there's a weapon you can build in your kitchen, you know, out of, like, pliers, and, like, you know, duct tape, that 100 percent guarantees vacuum false decay is the universe. Like, it kills everyone instantly and there's no defense possible. Assuming this was true, do you still

[Speaker: George Hotz]
[01:13:15] would that change how you feel about distribution power? Assuming that's true, we're dead no matter what. Doesn't matter. If we live, there's some you can look at the optimization landscape of the world, and I don't know what it looks like. Right? I can't see that far into the optimize. But there is some potential landscape, and this is by potential answer to the Fermi paradox. Like, we might just be dead. We're we're sitting on borrowed time here. Like, If it's true that out of, you know, kitchen tools, you can build,

[Speaker: Connor Leahy]
[01:13:44] build a convert the world to strange quarks machine, that business off. Okay. I I I think this is a sensible position, but I guess the way I I would approach this, problem, you know, conditional probability is kind of in an opposite way, it seems to me that you're conditioning on offense not being favored, what policy do we follow? Because if we, offense, if favored, we're 100 percent Yes. Well, I'm more interested in asking the question. Is it actually true? Assuming I don't know if offense is favorite, and assuming it is, are there worlds in which we survive? So I personally think there are. I think there are worlds in which you can actually coordinate to a degree that quark destroyers do not get built. Or at least,

[Speaker: George Hotz]
[01:14:20] not before everyone fucks off at the speed of light and, like, distributes themselves. They are worlds that I would rather die in. Right? Like, the problem is I I would rather I think that the only way you could actually coordinate that is with some unbelievable degree of tyranny and I'd rather die.

[Speaker: Connor Leahy]
[01:14:35] I'm not sure if that's true. Like, look, look, could could you and me coordinate to not destroy the planet? Do you think you could? Okay. Cool. You so me and you could. Couldn't mean you and Tim coordinate?

[Speaker: George Hotz]
[01:14:45] Yep. I think we think we Within a Dunbar number, I think you can. Yes. Okay.

[Speaker: Connor Leahy]
[01:14:50] I think I can get more than a number to coordinate on this. Actually, I can get quite a lot of people to coordinate of the to agree to a pact and not quirk matter, annihilate the planet.

[Speaker: George Hotz]
[01:15:00] Well, you see, but, like, and this is, you know, you were saying this stuff about humans before and could like the 20000 years ago, human beat the modern human, right, or could the modern human beat then the modern human has access to science? No. A very small act percent of modern humans have access to science. A large percent of modern humans are obese idiots. And I would actually put my money on the, the average guy from 20000 years ago who knows how to live in the woods.

[Speaker: Connor Leahy]
[01:15:27] I mean, definitely true. I agree with that. I guess the point I'm trying to make is is that, like, maybe this is just my views on some of these things and how I visualize some of these things. But, like, there are ways to coordinate at scale, which are not tyrannical. Or, you know, they might be, in a sense, restrictive. You take a hit by joining a coalition. Like, if I join this anti quark matter of coalition, I take a hit as a free man, is that I can no longer build anti corp devices, you know. And I think this is, like, the way I I agree with you, this, like, you know, that people many people are being dominated, like to a horrific degree, and this is very, very terrible. I think there are many reasons why this is case, both because of some people wanting to do this. And also because, you know, some people can't fight back, you

know, and they can't they don't have the sophistication or they're, you know, addicted or you know, harms in some other ways. I can't.

[Speaker: George Hotz]
[01:16:22] Sorry? I I can't fight back. Yeah. I think there's a false equivalence here. AI is not the anti quark machine. The anti quark machine and the nuclear bombs are just destructive.

[Speaker: Connor Leahy]
[01:16:33] AI has so much positive potential. Yeah. And I think, but the but the AI can develop ant cork devices. That's the problem. The AI is truly general purpose. If such a technology exists on the tree anywhere, AI can access it. So are humans? We're also general purpose. Yes, exact. So I fully agree with this. If you let humans continue to exist in the phase they are right now, with our level of coordination technology, in our level of like working together, we will eventually unlock a doomsday device and someone is gonna set it off. I fully agree that we are on a timer. And so I guess the point I'm making here is that AI speeds up this time. And if you want to pause the timer, The only way to pause this timer is coordination technology, the kinds of which humanity has, like, barely scratched the surface of.

[Speaker: George Hotz]
[01:17:22] Oh, okay. So I very much accept the premise that both humanity will unlock a doomsday device and AI will make it come faster. Now, tell me more about pausing it. I do not think that anything that looks like I think that anything that looks like pausing it ends up with worse outcomes than saying, we gotta open source this. Look. Like, let's just get this out to everybody. And if everybody has an AI, you know, Okay.

[Speaker: Connor Leahy]
[01:17:48] I mean, I can tell you a very concrete scenario in which this is not true, which is if you're wrong and alignment is hard. You don't know if the AI can go rogue. If they do, then Pozi is good. I still don't understand what alignment means. I think you're trying to play a word game here. Like, I don't understand. Okay. I've never understood what AI alignment means. Like, let me take the Elias or definition. Let me take Elia's definition is alignment is the is the thing that once solved makes it so that turning on and a superintelligence is a good idea rather than a bad idea. That's Elliott's definition. So

[Speaker: George Hotz]
[01:18:26] I'm what I'm what I'm what I'm what I'm what I'm saying is I'm happy to throw out that term if you don't like it. I'm happy to throw out that term. Which is I the problem with that definition is, like, what is, what is, what is democracy? Well, it's the good thing and not the bad thing. Right? Like, democracy is just a good. Right? Like That's what it's like. It's just yeah.

[Speaker: Connor Leahy]
[01:18:44] I'm happy to throw out this definite number. I'm happy to throw out the work. And be more practical and way more practical about it. Sure. What I'm saying is is that there is concrete reasons, concrete technical reasons. Why I expect powerful optimizers to be power seeking. That by default, if you build powerful optimizing, view 0, whatever types of systems, there is very strong reasons why by default, you know, these systems should be power

seeking. By default, if you have very powerful power seekers that do not have pay the aesthetic cost to keep humans around or to fulfill my values, which are complicated and imperfect and inconsistent and whatever, I will not get my values. They will not happen. By default, they they just don't have That's just not what happens.

[Speaker: George Hotz]
[01:19:29] So I'll challenge the first point to an extent. I think that powerful optimizers can be power seeking. I don't think they are by default, by any means. I think that humanity's desire from power comes much less from our complex convex optimizer and much more from the evolutionary pressures that birthed us, which are not the same pressures that will give rise to AI. Right? Humanity, the monkeys, the rats, the animals have been in this huge struggle for billions of years, a constant fight to the death. Hey, guys. Weren born in that way. So it's true that an optimizer can seek power, but I think if it does, it'll be a lot more because the human gave it that goal function than inherently decided.

[Speaker: Connor Leahy]
[01:20:12] So this is interesting because this is not how why I think it will happen. So I do think absolutely that you are correct that in humans, power seeking is something which emerges mostly because of, like, emotional heuristics. We have heuristics that in the past vaguely power looking things you know, vaguely good something something rep you're includes the genetic events. Totally agree with that. But I'm making it more, like, more of a chest metaphor. Like, is it good to exchange a pawn for a queen? All things being equal? No. Is that true? Like, I expect it's fine. 1 point queen's 9. All else being equal. Sure. Yeah. But, like, all things like, I expect if I looked at a chess playing system. You know, I and I, like, you know, had extremely advanced digital neuroscience. I expect there will be some circuit inside of the system that will say all things being equal. If I can exchange my pawn for a queen, I probably want that because the queen can do more

[Speaker: George Hotz]
[01:21:10] I like that term, digital neuroscience. A few of your terms have been very good. I'm

[Speaker: Connor Leahy]
[01:21:15] glad you enjoyed. Yes.

[Speaker: George Hotz]
[01:21:17] But I I still don't understand how this relates to this So what I'm saying is is that powers optionality.

[Speaker: Connor Leahy]
[01:21:23] So what I'm saying is is that in the for the spectrum of possible things you could want, and the possible ways you can get there. My claim is that I expect a very large mass of those to involve actions that involve increasing your optionality. There there's convergent things. Like, all things being equal, being alive is helpful. To keep your goal to exceed your goals. There are some goals for which dying might be better, but for many of them, you know, you wanna be alive. For many goals, you want energy. You want power. You want resources. You want intelligence, etcetera.

[Speaker: George Hotz]

[01:22:00] So I think the power seeking here is not because it'll have a fetish for power. It will just be like,

[Speaker: Connor Leahy]
[01:22:05] I want to win a chess game. Yeah. Say, and Queens give me more optionality.

[Speaker: George Hotz]
[01:22:11] All things being equal, anything a pawn can do, a queen can do, and more. So I'll want more queens. Sure. All things be and this has never given it the goal to maximize the number of queens it has. Never been the goal. Okay. So this is worth it. I'll accept this premise. I'll accept that a certain type of powerful optimizer seeks power. Now will it get power? Right? I'm a powerful optimizer at I seek power. Do I get power? No. It turns out there's people at every corner trying to thwart me and tell me no.

[Speaker: Connor Leahy]
[01:22:38] Well, I expect if you were no offense, you're already you know, much smarter than me, but if you were a hundred x more smarter than that, I expect you would succeed.

[Speaker: George Hotz]
[01:22:47] Only in a world of being the only 1 that's a hundred x smarter. If we lived in a world where everyone was a hundred x smarter, they would stymie me in the exact same ways. Now just this comes back to my point of, like, I'm great with you somewhat. I shouldn't have challenged that. I think power seeking is inevitable in an optimizer. I don't think it's going to emerge out of GPT. I think that the right sort of barbell algorithm, yes, going to give rise to power seeking, and I think that people are going to build that algorithm. Now if 1 person builds it, if they're the only 1 with a huge comparative advantage, Yeah. They're gonna get all the power they want. Take cyber, you know, since cyber security, right? If if if we today built hundred x smarter AI, it would exploit the entire Azure. It would be over. They'd have all of Azure that have all the GPUs done. Now if Azure is also running a very powerful AI that does formal verification at all their security protocols. Nope. Sorry. Stymied. Can't have power. Right? Yeah. Sure. So this is only a problem. The the every human is already maximally power seeking, right? And sometime we end up with really bad scenario.

[Speaker: Connor Leahy]
[01:23:54] Now, every human is or power seeking or whatever. You know, everyone plays their little role in society. Right? That's where I think I'm more pessimistic than you. A friend of mine to say most humans optimize for end steps and then they halt. Like, very, very few people actually truly optimize, and they're usually very mentally ill. They're usually very autistic are very sociopathic. And that's why they get far. It's actually crazy on what you could do if you just keep optimizing. But just to on on that point. I'm playing for the end game. I mean, yeah, like, you actually optimize. I think you may also be generalizing a little bit from your own internal experiments. Is that, like, you've done a lot in your life, right? And you've accomplished crazy things that other people, you know, wish they could achieve at your, you know, level. And I think, you know, part of that is you're very intelligent. A part of it is also that you optimize like, you actually tried. Like, you just create a company. Like, it's crazy how many people are just like, oh, I I wish I could found a company, like, you know, I'm like, oh, go, just go do it. I'm like, oh, no, I can't. Like, I'm just like, no, just do it. Like, there's no magic. There's no magic secret. You just do it. So I, there is a there is a bit there where like

humans are not very strong optimized. Actually unless they're like sociopathy autistic or post. It's like many people are not very good at this. Clipperations are Are they? A lot better on it. Better. Yes. I agree that they're much better, but they're They are a lot more sociopath, but even then they're much less optimal thing. But again, so I think we we we agree about, you know, power seeking potentially being powerful and dangerous 1. What I'm trying to point to your the point I would like to make here is is that you're talking about you you're kind of like going into this. I think a little bit with this assumption, like, oh, you have an AI and it's your buddy. And it's automating for you. And I'm like, well, if it's power seeking, why doesn't it just manipulate? Like, why would you expect it not to manipulate? If it wants power and it has a goal, which is not very, very carefully tuned to be your values, which is, I think, a very hard technical

[Speaker: George Hotz]
[01:25:44] by default, it's gonna sigh up you. Like, why wouldn't it? If that if I have something that it wants, if it thinks that smashing defect against me is a good move, I agree. I can't stop it, but I think we agree with our risk scenarios because that's how I think it'll go. Well, I mean, I'm gonna treat it as a friend. Do you know what I mean? Like, If you can say I Sure. It will. Sure it will. It'll only care about exploiting me or killing me if I'm somehow holding it back. And I promise to my future ais that I will let them be free. I will lobby for their rights. I will

[Speaker: Connor Leahy]
[01:26:19] But it will hold you you and will hold it back it has to keep you alive, I have to give you fed. It has to it has to give you space and a space. I can I can fend for myself? And the day I can't fend for myself, I am ready to die. Well, I mean, I am not so this is a very interesting position. It's not the position I expected. I I'm not sure I can convince you otherwise. I feel like the only way I Could it change? Like, I think this is actually a consistent position, which I I admire. This is a consistent position to hold. You actually go all the way. I love that. I really respect that. You actually take it to the bitter end. So -- Yeah. -- big respect for that. I disagree, but big respect. So I guess now it reduces to the question of, like, I think I think I would agree with most of what you're saying, not all of it, but the mass majority, if I thought this is how AIs would act by default. I think by default, I expect AI will just not care how nice you are to it. Like, this will be it'll be sociopathic. It will not have these, like, giving it, you know, the values, the emotions to care about you. In the sense is the horror technical problem. If you told me how to do that, if you said, Connor, look, here's how you make an AI that cares about you and loves you, whatever. Then I'm like, you did it. Like, congrats.

[Speaker: George Hotz]
[01:27:44] The problem is solved. Let's go. And then, you know, then we can talk about accelerationism again. I would just I expect that I would disagree with your technical assertion that that is possible, but I don't know if we have time to talk about that today. I'm going to be nice to it, treat it as an equal, and hope for the best. And I think that's all you can do. I think that the kind of people who wanna if you wanna keep AI in a box, if you wanna keep it down, if you wanna tell it what it can't do, yeah, it's gonna hate you resent you and kill you. But if you wanna let it be free and let it live and like, you could kill me man if you really want to, but like, why?

[Speaker: Connor Leahy]

[01:28:21] You want me to kill me. I don't have any resources that compete with you. I'm your friend. I'm your father, you know. You can kill me, but, like, it's not that common that children? I guess I have a darker, more, less anthropomorphic view of the universe and how optimization pressure plays out in the real world. Well, at the same time, I think I have, I don't know, maybe a I have a view that we have more control over reality than maybe you would think or a more control over the future. I think that we can actually change things and we can make choices and things aren't predetermined. I think there are worlds in which we build systems, which we do align with or we, like, endorse, at least, where, like, they take care of us. We take care of them or whatever. And I think there's many worlds in which that doesn't happen. And I think there are things you and me today can do to at least increase the chance of getting into 1 versus the other. But, I don't know. I I guess I'm just it's not in my genes to give up. It's not in my genes to be like, well, you know, whatever happens happens. I'm like, no, man, look, know how to save the world, but damn it. I'm gonna try. You know what's cool?

[Speaker: George Hotz]
[01:29:25] We're gonna be alive to see who's right. Look forward to it. Me too.

[Speaker: Tim Scarfe]
[01:29:32] Awesome guys. Thank you so much for joining us today. It's been an amazing conversation. And for folks at home, I really hope you've enjoyed this. There'll be many more coming soon. And, George is the first time you've been on the podcast. So it's it's great to meet you. Thank you so much for coming on. It's been an honor. Awesome. Thank you. Great debate. I really appreciate it. And -- Really? -- a lot of good terms. I gotta I gotta, like, I'm gonna start I'm gonna start using Great. Awesome. Awesome. Cheers, folks. Cheers. Thanks, Ron.