**Connor Leahy:** 0:00

What this whole open AI saga has shown us is that, I mean, obviously we can't have something like this being developed by, just like a handful of, you know, weird people, unaccountable billionaires in the Bay Area. This is actually something that I've been telling people for years now. Obviously, this is not the right governance structure. So the weirdest thing I would say about this technology should be a tool. It should not be a goal in and of itself.

**Craig Smith:** 0:24

Hi, I wanted to jump in and give a shout out to our sponsor, Netsuite by Oracle. I'm a journalist and getting a single source of truth is nearly impossible. If you're a business owner, having a single source of truth is critical to running your operations. If this is you, you should know these three numbers 36,025,1. 36,000 because that's the number of businesses that have upgraded to Netsuite by Oracle. Netsuite is the number one cloud financial system streamlining accounting, financial management, inventory, hr and more. 25 because Netsuite turns 25 this year. That's 25 years of helping businesses do more with less, close their books in days, not weeks, and drive down costs. One because your business is one of a kind, so you get a customised solution for all of your KPIs in one efficient system with one source of truth Manage risk, get reliable, forecast and improve margins Everything you need all in one place. As I said, I'm not the most organised person in the world and there's real power to having all of the information in one place to make better decisions. This is an unprecedented offer by Netsuite to make that possible Right now. Download Netsuite's popular KPI checklist, designed to give you consistently excellent performance, absolutely free, at Netsuitecom. Slash Eye on AI. That's Eye on AI, e-y-e-o-n-a-i all run together. Go to Netsuite,com slash Eye on AI to get your own KPI checklist. Again, that's Netsuite.com slash Eye on AI E-Y-E-O-N-A-I. They support us, so let's support them. Hi, my name is Craig Smith and this is Eye on AI. In this episode, I speak again with Connor Leahy. He's the founder and CEO of a startup called Conjecture that's working on AI alignment. Before that, he was one of the founders and leaders of a group called Eleuther.ai that built one of the world's first open source large language models, and Connor is concerned about AI safety, about where AI development is going, concerned about the push towards artificial general intelligence, and has a lot of thoughts about what we should be doing to control development so that we don't end up creating something that is harmful to humanity. I talked to him particularly because I wanted to hear his thoughts on the open AI saga, which highlighted for a lot of people the dangers of having such a small group of people controlling such a fundamentally powerful technology today. I hope you find the conversation as interesting as I did.

**Connor Leahy:** 4:08

So I'm Connor. I'm currently the CEO of Conjecture, an AI company in London focused on AI safety and building architectures for AI systems that are understandable and controllable and various other things. I also do a bit of work in policy, regulation, public messaging, that kind of stuff. Before this, I was well known as one of the founders of Alluther AI, which was kind of one of the first, if not the first, kind of like open source large language models, research, building kind of groups, and technically, even before that, I was someone who worked on open source GBT2 as, I think, maybe literally the first person.

**Craig Smith:** 4:52

Yeah, and so the last time we spoke was after the release of chat GPT and GPT 4. And you were very concerned, as were a lot of people, and a lot of people continue to be, about releasing these kinds of models into the public before having fully explored the safety issues or without having adequate guidelines. But it struck me during this open AI saga that we have lived through the last week that having this kind of powerful technology in the hands of a handful of people who have different agendas and can't get along is, in itself, a security issue. And I would think that having these models open source where everyone can see the data they were trained on in particular because even Llama, I've learned, doesn't make public the training data but having the data open for all to see and having the weights of the models open so people can improve them or play with them or whatever that to me seems like a much safer path than having proprietary models. But I wanted to hear how your thinking is involved in that.

**Connor Leahy:** 6:35

In the 1940s and in the early 1950s, the Soviet Union built what's called a closed city around what would be known as the Mayak facility. The Mayak facility was the largest and one of the first. Well, I don't know if it was literally the first, but it's one of the largest nuclear facilities for the Soviet Union. In their breakthrough development attempt to build the nuclear bomb To give a bit of a flavour for what Mayak and similar facilities that existed throughout the Soviet Union were like, there was a program where, if you got caught by the secret police and you were being sent to a gulag for the rest of your life, you would give them an option Either go to Siberia, work yourself to death for the rest of your life, or you could send to Mayak for only three months and if you serve your term, you're free. Sounds great, doesn't it? Well, no one survived the three months. So what happened, of course, is that one of the first things that the Americans developed while developing the bomb is the HEPA filter, which is a form of air filter that is powerful enough to be able to filter radioactive material very sufficiently out of the air, making it safe for the workers. The Soviets didn't bother developing HEPA filters? Not really, so a lot of people died. To this day, lake Karakai, which is near the Mayak facility, is one of the most radioactive places on earth, so much so that it is said that standing next to it for an hour can kill a man. These are all just some fun facts that really have nothing to do with the topic we're talking about today. It's a story about how some people who were pretty bad people developed something or were working on something pretty dangerous, and then other people got access to it and also really bad things out, because even worse people got access to it. Now, from this I don't conclude oh so we should have just had everyone develop plutonium. That would have made it safer. This is not the draw from this conclusion from the story. Now, nuclear power is obviously very different from AI in many factors, so why even bring it up? I can ask the question in the other direction, though. You're talking about AI, you're talking about AGI, so let's focus on AGI. I'm not really interested in talking about the risks of current day models like chat-chip-t or something. We can talk about those too. They are real. There are real risks from those, but they're not the kind where I think we have to stop all publication as a salary. But let's talk about AGI systems. What reference class does AGI fall into? Is it like open source? Is it like nuclear bombs? Is it like something different? The reference class we choose forms our thinking around something which is fundamentally none of those things. Agi is in nukes. Agi is open source. It's not Linux, it's something very different. So, while it has things in common with all of those things, agi runs on computers Linux runs on computers pretty similar you know and it has other things that come with nuclear bombs. Nuclear bombs can kill everybody. Agi can kill everybody. Those are things in common. Is it more like nukes? Is it more like open source? At some point we have to actually drop down from the metaphors and into the actual models of reality. So, from my perspective, I think you're completely correct. What this whole open AI saga has shown us is that, I mean, obviously we can't have something like this being developed by, just like, a handful of you know, you know weird people, unaccountable billionaires in the Bay Area, like, obviously they're not acting in humanity's best interest. No one's surprise, like a couple of months ago, Sam Altman was, you know, interviewed about this and he said, oh yeah, the board can fire me at any time If I go by the mission. I think that's important. And then they try to fire him and he's back and well, that didn't work. So this is actually something that I've been telling people for years now. I've gotten some, you know, disagreements he did disagreements, let us call them, with some of the people who were involved with the creation of the board and, or you know, in favour of the existence of the board, and the point I always made to them was just like this obviously cannot control a you know charismatic billionaire, you know political mastermind. Like why the hell would you think it would? This is crazy, this is, and this is exactly what we saw play out. And I'm not even trying to make a comment on Sam Altman good, Sam Altman bad. I'm just saying obviously he was just not going to just say, oh gosh, darn. I guess the board said no more AI for me. Guess I'm going to stop. That's not how men like him work and that's obviously not what was going to happen. And you think, like the politically unsavvy nerds you know, could write a document that would convince someone like him to stop? No, of course not. So obviously this is not the right governance structure. I fully agree with this. But there's a great saying, which is that reverse stupidity is not intelligence. If you're taking something stupid and you take the opposite of it, it's probably also stupid, and so this is so. The fact that this governance structure doesn't work for me does not say that, therefore, there should be no governance structure. This, to me, does not follow.

**Craig Smith:** 12:00

Yeah, so, but open source. I mean you are, I'm sure, very familiar with Yanlacun's argument that, yeah, that open sourcing can lead to some abuse by bad actors, but by and large, the vast majority of people that will be working on an open source model, contributing to it or building products off of it, will be doing so with you know, not without nefarious intent, and that the larger the open source community, the quicker it would be able to respond to bad actors or misuse, or the more people available to build guardrails and spot weaknesses and that sort of thing. So that argument makes a lot of sense to me. I mean, at the beginning, when the, you know, pause letter came out and around the time that we talked, I thought, yeah, this stuff is too dangerous to be open source. But I'm changing my mind and I wanted to hear whether this episode has changed your mind at all.

**Connor Leahy:** 13:39

I'll make three points in reaction to that. The first one is a story, the second one is a heuristic and the third one is a true observation from my own life. So first the story. The story is that the smallpox virus genome is currently online. You can go download it. It's a small text file, you can just go download it to your computer. Another fact of the story is that a couple of years ago Canadian scientists recreated an extinct version of smallpox called horsepox. They revived it they may and it was functional and viable and infectious and they published how to do it. Do you think either of those things are good? Now you can argue. Well, if we have more eyes on the smallpox virus, then something, something you know, good things happen. But this isn't really a model. So this brings me to the second point. The second point is offence versus defence. The way technologies work is that some favourite offence, some favourite defence very few are symmetric Most of the time, and most of the time, offence wins. It is usually easier to destroy than it is to protect. There are exceptions to this rule. For example, cartography is an interesting exception where defence is easier than offence, but in most cases it is easier to build a bomb than it is to build a reactor. You know a safe, controlled burn. So, all things being equal, you should expect that if you have a technology and you distribute equally, that there will be more destruction. This is the default case. This is what you should expect by default. Most technologies that are destroyed don't immediately give you a way to defend against it. Developing vaccines is harder than developing bio weapons. It is much easier to crank out a bunch of bio weapons and then you have to develop vaccines in response to that, which is already super hard, because who knows how far the virus already is. So just because the technology is why the aspect does not mean that defence wins Whether offence or defence wins is a property of reality. It is not a property of your morals or of your ideology. And the third point is an observation from my own life. I used to work in open source. I was one of the very first people to work on it and I had similar views to Licken and you know, assuming he holds these views genuinely, which you know, I hope he does. I don't know him very well, I've talked to him maybe once, and I think this is just not even wrong. It's just in my experience, what happens when you build AI models and you release them open source is that the first thing that happens is to get uploaded to a hugging face. And then a guy called the bloke that's literally his name, uncensored system, undos, any RLHF training or other security training run, trains them with all the newest data to make them more powerful, more general, more whatever uploads them again. 4chan, downloads it down, you know, uses them for whatever their applications are, whether it's, you know, pornography mostly, or likes BAM or whatever, etc. And now maybe this is fine, right, like you know, you know, maybe we say it's okay if people want to use their LLMs for porn, so what, that's okay, sure, what I'm saying is is the empirical observation is that the amount of effort that gets put into making these things safer or more controllable is absolutely pathetic Compared to the amount of effort that the open source community puts into making these things more powerful, more general and less controllable. This is just an empirical fact. This is just. Actually, if you go online, you pick the top 1000 LM repos, how many of them are about controlling the models better versus making them faster, making them more efficient, distilling them, making them more, etc. And the fact is that the offence, like the unbalance here, is like it's not even funny and I understand right, and this is not to say that the people working on this technology are morally evil. I think this is an important thing to understand. There's an incentive from people like Lacan and others like big tech, you know, people like talking heads to try to focus on it's only the evil people's fault because that absolves them of responsibility. Meta wants open source because it absolves them of responsibility as a corporation. They can't get sued because it was the user's fault. This is also what's happening in the EU AI Act right now is that people like Lacan are lobbying to remove foundation models from regulation in the EU and saying it says their uses should be regulated. This is the same thing as when, for example, plastic companies invented recycling. They invented it so that it was the user's fault that there is all this plastic pollution. Like oh see, we would have recycled it, but unfortunately the users just didn't do it. This is gaslighting and this is a complete unbalance of power. The externalities of plastic pollution should be on the ones who are most suited to addressing this externality, who are creating this externality. It shouldn't be on the user. The same thing applies to foundation models is that these systems can do things. They can be used for many things and we should be taking the big companies building these systems. It's not like these open source models are being built by plucky little teenagers in their rooms. As a plucky teenager that did do that, I'm saying most of the ones being built now are being made by the UAE and Metta. These aren't the little guys, they're big guys trying to shirk their responsibility to society.

**Craig Smith:** 19:21

Well then, what's the lesson from the open AI saga? That you just need a bigger board.

**Connor Leahy:** 19:31

The lesson is that none of these structures are correct. This is what we have governments for. This is the same lesson that we've had over and over again: Self-regulation does not work. It has never worked. This is like tobacco companies self-regulating themselves. This does not work. We, as a society, have developed a mechanism. I'm not saying it's a perfect mechanism by any means, but we do have a mechanism for intervening in systems that have extremely high externalities that are not self-regulatable, and it's called the government.

**Craig Smith:** 20:04

Yeah, I mean there has been a lot of work at the government level, not as much in the US as in Europe. But how are you? I mean, obviously these models are so commercial the potential is so commercially exciting that fines aren't going to matter. You're not going to be able to find people to behave in ways that the government wants them to. There's got to be something stronger than that. So, Jeff, you thought about that. I mean, how do you regulate these things?

**Connor Leahy:** 20:52

One of the most inspiring moments from the history, I think, of science and society is that many decades ago, biologists and chemists and so on realised that human cloning should be possible. It should be possible to do this. They were still far from having the actual technology to perform human cloning, but they found out it should be possible and they reasonably understood. Wait, that might be really disabilizing. That could be. We don't know what the consequence is. Maybe it's great, maybe there's many benefits from human cloning as well. But let's chill out. We don't know. We don't know. And this seems huge. This isn't just another thing. This is not a 10% more effective cough drop. Human cloning is a big deal and so, heroically, long before the technology existed, they came together and banned it and said let's have a moratorium, let's not do this until we've had a bit more time to figure out what the hell we as a society want about this. It wasn't one board. This wasn't one CEO being like I will take a moratorium on this. No, it was the scientific community and governments coming together and working very, very hard to create a moratorium. A moratorium is what we do when we are faced with something which we know is huge and we don't know how to deal with. That's what scientists do. We have a moratorium and we should have a moratorium on AGI. This is what we need to do. Can you enforce a moratorium? Yeah, I mean it's technically, like physically, like, yeah, obviously, like that's not that hard, whether people will do that, whether people want to do that, whether people can overcome the incredible political power that Big Tech has. That's the more interesting question. It's not like that. The government obviously has the ability, like the CIA, can track every GPU in the country if it wants to. If the NSA wants to shut down, just press a button. That's not the problem. If you want to throw a couple CEOs in jail, sure, like the FBI can do that. Like physically, this is not a problem, it's a political problem. This is not a physical problem, this is a political problem. The political problem is well, if you have legislation around this kind of stuff well, we just saw what happens. If you try to fire Sam Altman you think he's going to be okay with taking his GPUs away. Well, no, I expect that's going to be a hard fight. I expect Microsoft lobbyists will fight that tooth and nail. I expect many people will fight this and this is why, like you know, I'm not here to point you out, paint you a rosy picture of the future. I'm not optimistic that things are going to go well. We have an unprecedentedly huge political problem here. I think I'd like to say that the thing that's killing us right now is not AGI. Agi doesn't exist yet. It's people, it's politics that is killing us.

**Craig Smith:** 23:46

Right, right, but to that point that AGI doesn't exist, not so much all the other, I mean. Yes, no doubt the political systems are not equipped to deal with the big problems facing humanity, but in this case AGI doesn't exist. I don't know how you would ban AGI, because no one really knows how and when it might emerge, if it ever does, at the level of the tech now, I mean, what are you suggesting? And I'm not putting it on the spot, I don't expect you to have, oh, I have policy proposals.

**Connor Leahy:** 24:30

I have very concrete policy proposals. Here are three. The first one is a computer cap. There should be a limitation that no single training around, no single AI system can be built with more than a certain amount of compute. So luckily we are very lucky that current frontier AI systems, more and more general purpose systems, require more and more computing resources. These computing resources are very easy to track. They're very bulky. They take lots of specialised knowledge, lots of energy. The kinds of supercomputers that can train a GPT-4 or a GPT-5 are only built by like three companies in the world and they're all in the US. So this is a solvable problem and we should put a ban on it. You know, there should be a registration process for frontier models up to a certain limit and beyond that there should be just a moratorium. You are not allowed to perform any experiment that requires more than 10 to the 24 or 10 to 25 or whatever flops, Flop being a unit of measurement for computing power, and this is easily enforceable. This is absolutely something that technically is enforceable. It's just a political problem and this buys you time. When our scientists figure out, you spend time actually figuring out how far is AGI away. How dangerous is it? How do we control things? Blah, blah, blah. Then we can talk about those kinds of things. The first thing is to buy time. The second proposal, unless you want to comment on that.

**Craig Smith:** 25:57

Well, just on that. You're talking about limiting commercial products. But when you say, then that gives the research community time to figure these things out, they're going to have to experiment with larger models, so there's got to be some.

**Connor Leahy:** 26:23

To be clear, these levels are insane 10 to the 24, 10 to the 25, flop is an unimaginably large amount of computing power. There are no academic labs, basically, that need this for research, FD research. This is ridiculous. There is just no. So this is a common propaganda piece that big labs like to say is like oh, we need more computers to do safety research. Maybe this is true. I have not seen it. This is just not what has actually happened, Just purely empirically speaking. There is, I have seen, basically no safety AGI relevant research that required more than like a GPT-3 that you couldn't have done with GPT-3 level of compute or less. Maybe it exists, but I sure as hell have not seen it.

**Craig Smith:** 27:16

Limiting computers is one proposal. What are the others you mentioned?

**Connor Leahy:** 27:23

To others. I would recommend the second is strict liability for model developers. So what this means is strict liability means that the intentions of the developer do not matter. What matters is that if harm is caused, the developer is liable. I think this should basically exist for the whole supply chain is that if you create externalities, you have to pay for them. This aligns the incentives of everyone aligned on the chain. Currently, there are no incentives for developers to minimise the externalities of their systems. Currently, you, as an open source developer, can be an arbitrarily dangerous thing that causes arbitrarily much damage and you have no incentive to avoid this. As a concrete example, which is not even going to AGI, is voice cloning systems. There are right now in GitHub systems you can just download which can take 15 seconds of your voice, clone it perfectly and call your kids, call your wife, manipulate them, call in a swat, hit on you using your own voice. This is all doable and the people developing these systems have zero liability. They don't even feel bad about it. They don't even have any conscience, because it's open source. Craig, if it's open source, it must be good. My ideology says so and you know, when your ideology tells you something is morally right, then it's good, as we've seen throughout history. So we have to align incentives here somewhere along the line. It reminds me of cars and seatbelts in the 70s, where car manufacturers fought tooth and nail to not have seatbelts. They fought it viciously, with propaganda and with lawsuits and with everything they could throw at it, because they said well, it's the driver's fault if he gets into an accident, it's not our fault. We just build cars. If they drive it poorly and they die, well, it's not our fault. And we, you know the people rightfully told them to go fuck themselves, like. No, you have to build a safe product you can't like. It's not a moral question. It's kind of like the point I want to make. I'm not making an ideological point, I'm not saying my religion says that seatbelts are good. I'm like I don't care, I care. Do seatbelts mean that less people die? And the answer is yeah, like they make cars safer. So then I want seat belts cool. And the same thing applies to open source. Does Linux being open source result in more safety? The truth is, yeah, it looks pretty obvious like a case. So I'm in favour of Linux being open source. Awesome, great, you know. Do you know some seven billion parameter models? Be open source positive or negative? I don't know, probably positive, like probably so, I'm not sure. Like there's a lot of downsides there as well, but like, seems like it probably is positive, agi being positive, you know, open source, you know that does not seem positive to me at all. That does not. That seems like a recipe for disaster. So it's. I'm not trying to make an ideological point. That's what I'm starting to say. I'm not saying all these things are good, all these things are bad. I'm saying we have to look at things on a case by case basis. This is how proper regulation works. Proper regulation shouldn't be ideological. It shouldn't be that everything is regulated as ARB. That would be terrible regulation.

**Craig Smith:** 30:47

Yeah Well, so that was the capping of the computer on training runs, shifting liability to the model developer. What was the third one?

**Connor Leahy:** 31:02

So the third one that I think should be done is that there should be a kill switch. Now, what I mean by this is it doesn't have to be literally a switch. What I mean is there should be a protocol that any developer of Frontier AI systems needs to implement, by which, at a given notice, any Frontier training runs or deployments can be shut down in under a minute. So the reason for this is not per se, because I think that this would be very helpful. The AGI actually happens. If AGI actually happens, this is probably useless. The reason I think this is good is because we should have the institutional capacity to do these kinds of things. There should be, every six months, there should be a fire alarm, there should be a fire drill where everyone has to practise. Like in the next five minutes, all AI companies have to go offline for 60 seconds, if not, you'll get slapped with a huge fine. These are the kinds of protocols you want to have in worlds where you have tail risks, where things can blow up, where you can have these kinds of things. And then there should be a multilateral, K or N kind of system around this, like maybe all major global powers have one of these buttons and if three or five of them push it, or like seven or 10 or whatever, then the system kicks in. This is the kind of institutional building which doesn't save us, but it's a hell of a lot better than nothing.

**Craig Smith:** 32:25

And how do you see these kinds of proposals moving through the policy making frameworks? I mean, there is some advance in the European Union. You know, the White House has come out with its executive order, which as yet doesn't have any real concrete government governance policy in it, but it sort of lays out the things that we should be thinking about. Yeah, where do you see these things going? What sort of a timeline do you think that governments are being educated enough that they can deal with this? Who, what government is gonna lead? Is it the EU? Will it be the US? Who should it be? And then of course, you've got the other side of the world, Russia and China, who have very different agendas and may not wanna regulate at all.

**Connor Leahy:** 33:39

So when people ask me questions like this and they're like what's your probability of X happening? And then my follow-up question is usually is it X conditioned on me and other people doing something about it or not? Cause I expect if a condition on me and other people don't do anything about it, then yeah, I just think nothing will happen and big tech wins and then we die. I think it will be very heroic or special. It will just be that new products keep happening, AI keeps going up, and then just one day humanity's not in control anymore and we have no idea what's going on, and then it's just over. I don't think it will be dramatic. I think we will just get more and more confused. We won't understand what's going on anymore. Weirder and weirder things will happen. More and more politics, economics, markets, media is controlled by AI or even just fully generated by AI. There will be no more movies or just AI generated, and then just humanity will not be in control anymore. And then one day we fall over dead for some reason we don't understand. That's why expectations will happen by default and along the way. To be clear, big tech will be a lot of money. So go buy that Microsoft stock. You'll get really rich just before you die. So if I could add on someone actually doing something about this, I do think there is hope. I don't think there's a lot of hope, but there is hope, and the main hope I see from this is is that the general public fucking hates AI. It's unfathomable how much normal people hate AI. They use it, of course, but they're freaked out by it, which is just completely the correct reaction. It's just these crazy, bizarre, weirdo tech people like you and me who don't instantly like to wait, that's actually like, let's not do that. If you talk to any normal person, you're like hey, these people are building systems that are smarter than humans. They're like don't do that. That's no, that seems really dangerous, don't do that. Well, all the types of people are like oh, but actually, you see my proposal because we'll make it fine, or actually, universal love means that AI systems will love, or whatever. I don't even know what these people say anymore. I think they've given up making arguments at this point and they're just vibing, so I don't even know if there's an argument that debunked there. So from my perspective, we are building systems. They are going to be built by default unless we do something about it. So the general public, once these systems are not built, or at least for us to slow down until we can make them safe and we understand them better and they've been integrated into society, et cetera, et cetera. So now you might ask the question okay, well, that's true, why is fuck all happening? And that's a good question. And now we have to talk about models of policy change and like global coordination, which, at least how I think about this problem generally, is that the general public actually does have power in the West and like in democratic countries that it's very fashionable among elites to sneer and be like oh, actually, you see, the populace, they don't have true control. We live in, whatever the words are that people like to use, and this is to a large degree true, but it's not fully true. The main problem is that the general public has extremely short attention spans and is extremely discoordinated. This is the main problem. The bottleneck on policy action currently is not will of the people, it's not ability to enforce regulation, it's coordination. It's getting people to actually do something about it, to actually write letters to their senators, actually put things on their desks, actually yell at them on the phone, actually like to talk about it on social media, et cetera, et cetera. This is the kind of thing that's currently missing. Basically, campaigning, this is the kind of stuff that is missing, and I expect that if you did this well, if you raise this to saliency about people, you wouldn't have to convince them and I'm saying this because empirically, this has been true in my experience. Like talking to people and also doing stuff like focus groups and stuff, I found that you don't really need to convince people very much. You mostly just have to tell them facts, just have to, you know, just present them with hey, this is what's going on right now, and then mostly they converge to the reasonable beliefs around like, hey, that's scary, don't do that. So I think this is currently the best path we have. I'm also, you know, excited to talk to politicians and I talk to many of them, mostly in the UK and the EU because I'm UK based. But it's hard Because you know politicians have similar problems. They have very little attention span because they have so many things they need to do, there's so many things haranguing them, and my model of policymakers is basically that the ultimate goal of a politician is to not get blamed. So it's because, as a politician, you really have like I have. So if there's any policymakers listening or any staffers or so on, I feel you, I, you're in a shit spot. I get it Because, like, basically, the way I see it is like there's kind of like a two by two grid of like what you do as a politician, what you can do. So the idea is that there's a default action. Is that in a common, in our common you know feelings around an issue, there's something that is the default thing to do, which is usually nothing. If you do the default action and it goes wrong, well, you're not blamed, you know, because you know you did the sensible thing, not your fault. If you do the default action and it goes well, well, great, you're a genius, you know a good job. If you do the non-default action and it goes great, cool, yeah, you're good, great. If you do the non-default action and it goes bad, then you get blamed. That's how you get blamed. So you may notice from this payoff matrix that it is always better to take the default action rather than a non-default action. It is always better for the politician to not stray off the path, and this is universally true. So it's easy to yell at politicians and be like they have no spine, they have no courage and whatever. And yeah, that's true for many of them. Many of them are just yeah, just, you know. They just don't care, true, but some do, and they do go off the path and they get burned for it, and that sucks, but it is how the game is. So what we can do as the people is, we have to change what the default action is. You have to change the narrative from I guess we just keep bumbling along until we die to how the fuck dare you keep bumbling Like? Seize your bumbling immediately. Bumbling is no longer accepted, and that's my biggest hope at the moment.

**Craig Smith:** 41:00

Yeah, when we spoke last time again right as GPT-4 was being released, one of your immediate concerns was that these things can be hooked up to systems that can take action, and I don't remember if we talked about auto GPT, that first. I haven't looked at what's happened with that, but that first attempt to create an agent that could use LLMs, but that has developed a pace and we're now on the cusp of seeing sort of an explosion of AI agents that can leverage the power of large language models or other tools. I had a guy on earlier from NewsGuard, a company that builds databases to try and help companies, tech companies, identify disinformation and combat it, and we were talking about, once you have these agents, building, you know, creating disinformation not only creating the disinformation, but distributing it on a massive scale and maybe on a massively parallel scale, the internet, public discourse everything is gonna get very confusing because you're not gonna be able to tell what's real and what's not real and people which is the majority who are not particularly careful about where they're getting their information, will be manipulated. So, yeah, the coming AI agent era. How do you deal with that?

**Connor Leahy:** 43:16

I mean, I don't know, get your affairs in order. A number of years ago, after post GPT-2, this was around GPT-3 time, that's how we marked the eras. Now it's like instead of years, we just use GPTs. Now I was invited to work kind of like a discussion group with some open AI people, policy people like disinformation experts and stuff like this about the potential for misinformation and so on from language models. So it's like before GPT-4, before chat GPT and so on. So it was polite to lead all these like well-credentialed experts with their triple Stanford professorships or Harvard whatever talk about like misinformation, bias and whatever. And then when it came my turn to talk, my reaction was like, holy shit, you're all so undressed like you're being so optimistic. It's so much worse than any of you. You're like, oh, it could make it easier for far writers to just I like that's fucking children's play compared to what you could do with these things. Like you were truly. You're not creative. Like if you think that's a little worse, that can happen. Oh, they're gonna generate some fake news and some like Russian dissid fo websites. I mean, oh boy, that would be nice. That's a nice timeline. It's gonna be much worse than that. It's already getting worse. Like that. Talk about fully automated cults with fully automated profits. Talk about full, all sensory illusionary interactive systems creating full complex narratives that are completely disconnected from reality. Talk about full epistemic collapse, the semantic apocalypse. Even if AI's don't kill us, they're gonna drive us insane. So it's because it will just be harder and harder and harder to survive in a more and more adversarial informational environment. This has already been happening for a very long time. We just had Thanksgiving and, as much as we love her, we all have that one aunt that get way too into QAnon a while back and imagine so currently, stuff like QAnon or like I don't even know if QAnon's still a thing, but like whatever the newest thing is, the newest cult is newest, whatever is that affects some percentage of the population, some percentage of the more vulnerable population I can say stupid, just like, maybe emotionally vulnerable or epistemically vulnerable and for some reason not trying to judge these people here. Now imagine the bar keeps rising. You get systems that become more and more convincing, that become more and more sophisticated, more and more targeted, and slowly, slowly, the number of people who are just functionally schizophrenic keeps going up until at some point people cannot converge on reality anymore and just every person you meet is functionally schizophrenic. You cannot run a society, you cannot organise a system, if you and your neighbour cannot come to a conclusion about basic reality. This is like what is possible with these kinds of systems. I'm not saying this is gonna happen next year. I mean maybe, but these are the kinds of things you can do. Epistemics are hard. Like this is the thing that. Like there's also things like honesty is hard. This is like some people are like oh, just misinformation is a trivial concept. It's almost become a slur at this point. It's become a joke when people use the word answer information. Like at least in my social circles, a lot of people rolled their eyes to be like, oh yeah, anything that isn't big media, isn't this information, whatever. But it's just not that easy. Like finding out what is true, and disseminating and evaluating what is true is hard. This is very hard. It takes energy, it takes effort, it takes mechanisms, it takes like it's hard and it's going to get harder, it's going to get more expensive. The current light, like do you really know what's happening in Ukraine right now? Really I don't. I think I'm at a point where it is literally impossible for me to actually know what's going on in Ukraine. It's something that affects me, you know, affects family friends. You know it is a huge thing. I don't think that there is any way I could actually acquire and verify the truth of what is actually going on there. And this generalises this even before we get into agents doing worse things than this. I mean automating all jobs. Obviously, you know anything you can do at a computer an agent will do better and faster. So there will be complete economic collapse from that, like, obviously there will be no more need for human jobs until the inference costs, you know, get too high, but you know you can improve those back down. You'll have systems that can do harm in various ways. You know, by manipulating markets, campaigns, politics, you're gonna have systems that are, you know, cyber crime, hacking. You have a system Like it's like. When you ask a question like what is the worst thing agent-based systems are doing, you're asking the question what are the worst intelligence systems can do? What is the worst that a human can do? The answer is a lot.

**Craig Smith:** 48:47

Yeah, but again, yeah, I mean you can see that very bleak future, but I'm also a great believer in how mankind, the worst case scenario generally is not what happens and people kind of muddle along and-.

**Connor Leahy:** 49:26

But that survivorship bias. There was a man named Stanislav Petrov who was a Russian soldier stationed in Nuhuya bunker and he had the command that if American missiles appeared on the screen, he shot the missile. And one day six missiles appeared on his screen. His commands were very clear. The second guy with him there, who had, you know, the other key, was ready to turn and yelled at him that it's time we had to shoot back. The Americans were attacking, and Stanislav didn't. He disobeyed orders. He could have been, you know, fucking executed for that. And he disobeyed orders that day. And it's because of this one man, one Russian soldier, that you and I weren't nuked. One guy, we got lucky. So when people say, oh, but so far as I have, like what the fuck are you talking about? This is like saying, well, I've played Russian roulette five times so far and it's been great, let me pull again. That's just not how anything works. This is not how reality works. If you play like this and then eventually you predictably loop into it, you predictably lose. You have to play strategies where you can win in adversarial environments, where you can play where you can win in games where dangers exist. Our ancestors, when they were in the wild, couldn't be like them. Well, oh, my forefathers survived, so I don't have to worry about bears. You know, none of my forefathers got killed by bears. No, like that's just no. This is not how things work. The world isn't nice. There is no arc of history. There is no war that is protecting us. The fact that we are here today is because of the hard work of our ancestors. The fact that I live in this nice, you know, warm apartment sounds like, safe, that I have enough food to eat and so on, is not God that gave me that. It's not some, you know, force of nature. It was the hardscrabble and bloody fight of my ancestors that left me this. And if I let this rot, if me and other people don't maintain society, then it just dies, like then, entropy wins, entropy always increases and entropy is death. So if we just sit back and hope things will go well, they will not.

**Craig Smith:** 51:40

So you know I was gonna. I was thinking, well, that's a good place to end it, but I don't want to answer there, because our last conversation got an inordinate number of views and I have some producers that take these and turn them into shorts and they have these sound bites from that episode that have gotten an enormous number of views Because people gravitate towards these doomsday proclamations and I don't I mean whether or not they're true. I want to end on something more hopeful. So what should people do in your view? What should regulators be doing? What should researchers be doing?

**Connor Leahy:** 52:40

What should I do?

**Craig Smith:** 52:41

Microsoft is doing it.

**Connor Leahy:** 52:43

So the weirdest thing I would tell you. It's clear. I don't like being the doom guy. I absolutely don't like this. I was the pecto optimist throughout my entire life. I was always the person saying no, we can fix problems. Climate change is solvable. Solar powers can be exponentially cheaper. We can do carbon capture. There are so many things we can do. I've always been saying no, see how, in the interest of improved education, how much people are becoming better at having more access to information? Look at how so many things are. I was just reading the other day about how slowly over decades, just the flash freezing of frozen food has gotten better. And I've noticed that it's like my frozen broccoli. I'll make it night. It's just a little bit nicer and you know what. That might sound like a teeny thing compared to all these other things, but I think that's beautiful. I think it's extremely beautiful that life gets better. All things being equal, life has gotten a lot better. I'm very happy to be alive right now. All these small things done by these smart people, mostly done for profit Sure, the broccoli company, they just want profit, but ultimately they made my dinner a little bit nicer. It was already fine, I was already surviving, but it was a little bit nicer, and you know what. That's awesome and it's so nice that we can live this way. The truth is that we are so lucky that we live in a society full of educated, smart people that, for the most part, not all of them are angels. They're not heroes, but they do want to leave the world better. They want people to be happy, they want people to be safe and, all things being equal, almost everyone is given the option. If they could just help someone else and it didn't cost them anything, they'd do it and that's really nice. So we have to leverage this. We have to leverage that we and this is not the case everywhere in the world I want to say this is something that even today, is not in every country, it is not in every place or in every society, but in the West and many other countries in the Far East and so on, most people are educated. Most people are decent. Again, I'm saying they're great or heroes, but they're decent and they want the world to go well. They want their kids to grow up and have a nice life and eat nice frozen broccoli, whatever. They want to see art and beauty and music and so on, and we can have this. This is the important thing to understand. The important thing is Sometimes I'll talk about this idea of techno-optimism, quote-unquote. It's just cynicism in disguise. This is a really important thing to understand. These people who talk about, oh yeah, actually we're techno-optimists, we're accelerationism or whatever. They're just cynics. They're just libertarian cynics that don't believe that society can be improved except by just giving themselves to this abstract process of technology. But technology is not a force of nature. It's not a thing happening to us. It's a thing that we do. It's about humanity. It's not about technology. Sure, technology is great, it's helped humans. But I only care about technology because I care about humans, I care about people and we all care about people. We care about our families, we care about our friends, and technology should be a tool. It should not be a goal in and of itself. So when people talk about, well, agi is inevitable, someone's going to do it. No, no, it is not. It is not inevitable. It is not a force of nature. It's a decision we make. It is a decision we make and we can do better. We can, as people, societies, as civilizations, make choices. We can say, hey, let's be a little more careful. That doesn't mean we'll not do any AGI anymore. We can just say, hey, give our scientists a couple more years, a couple more decades, to understand the mathematics of interpretability better, and then maybe we'll give it another shot, you know, like we did with human cloning. These are what's important. I'm not saying that this is easy or that this is what's going to happen. It's not what's going to happen by default, but it's just important that there is this poison in our society that believes that the future is already decided. And it is not. The future is not yet decided. We still have a choice. It is not yet too late, but it will be soon.

**Craig Smith:** 57:08

Hi, I wanted to jump in and give a shout out to our sponsor, Netsuite by Oracle. I'm a journalist and getting a single source of truth is nearly impossible. If you're a business owner, having a single source of truth is critical to running your operations. If this is you, you should know these three numbers 36,000, 25, 1. 36,000 because that's the number of businesses that have upgraded to Netsuite by Oracle. Netsuite is the number one cloud financial system streamlining accounting, financial management, inventory, hr. And more. 25 because Netsuite turns 25 this year. That's 25 years of helping businesses do more with less, close their books in days, not weeks, and drive down costs. One because your business is one of a kind, so you get a customised solution for all of your KPIs in one efficient system with one source of truth: Manage risk, get reliable, forecast and improve margins Everything you need all in one place. As I said, I'm not the most organised person in the world, and there's real power to having all of the information in one place to make better decisions. This is an unprecedented offer by Netsuite to make that possible Right now, download Netsuite's popular KPI checklist, designed to give you consistently excellent performance, absolutely free at Netsuite.com. That's I on AI, e-y-e-o-n-a-i all run together. Go to Netsuite.com To get your own KPI checklist Again. That's Netsuite.com slash. I am on AI E-Y-E-O-N-A-I. They support us, so let's support them. That's it for this episode. I want to thank Connor for his time. If you want to read a transcript of the conversation, you can find one on our website. I am AI. That's E-Y-E-O-N-A-I. As I always say, the singularity may not be near, but AI is changing your world, changing it rapidly, so pay attention.