**Yoshua Bengio Guest** 00:00

We know how to build a superhuman AI, and if that recipe is publicly available and, even worse, if the weights of a trained AI are available, then it becomes very easy for a huge number of people to issue instructions that, if they were executed by the AI, could lead terrible catastrophes and, in the worst case, complete extinction. It's very hard, in terms of your feeling of your ego and how you can feel good about what you do, to accept the idea that what you've been working on for decades might be actually very dangerous for humanity.

**Craig Smith Host** 00:42

In a 1955 paper titled The Computer in the Brain, John von Neumann wrote that the possibility of replacing the human brain by a machine which will be superior to it in any or all respects is not excluded by any natural law that we know. It's therefore possible that the human race may be extinguished by machines. Von Neumann's warning has since been echoed by many other AI experts, from Stephen Hawking to Elon Musk. The concern captured the public imagination earlier this year when the Future of Life Institute issued a letter calling for a six-month moratorium on large language model development, saying that the development of artificial general intelligence could lead to human extinction. Then, in May, the Center for AI Safety issued a statement warning that the development of artificial general intelligence could pose an existential risk to humanity.

01:52

Many who had not signed the Future of Life letter signed this one, including Jeff Hinton and open AI CEO, sam Altman. Joshua Benio, who signed the Future of Life letter, had spoken on this podcast before about the threat, but I spoke to him again specifically about the extinction threat, whether it's a useful characterization or simply adding anxiety to a generation already troubled by looming existential threats such as climate change. Sir, just a moment, sir. Yes, I apologise for interrupting you, but I wanted to remind you that Oracle is offering a full implementation of.

02:38

NetSuite. Thanks for reminding me. Please continue. Yeah, okay, you can go back to your station. Thanks, yeah, beauregard reminds me that for the first time in NetSuite's 22 years as the number one cloud financial system, you can defer payments on a full implementation for six months. That's no payment, no interest, for six months.

03:06

To take advantage of this unprecedented offer, go to netsuite.com backslash i on AI, that's www netsuite com, backslash i on AI E-Y-E-O-N-A-I. I'll run together. Now here's Yashua. Hey, yashua, i wanted to talk to you about your essay on how an AI could become rogue, and since then this extinction debate has ramped up. On my Twitter feed there was a long thread by somebody with a blue check mark. That doesn't mean much anymore, but basically bemoaning the generation we've already destroyed, because extinction is inevitable now that we've developed these language models. And that's just me. But it just to me that debate, which is bleeding into the general public discourse, has gotten out of hand. But your essay was really the first concrete response I've seen on how things could go wrong. So I wanted first for you to just present that scenario And then I'll ask some questions.

**Yoshua Bengio Guest** 04:47

I don't know if I told you last time, but I'm also writing another blog post which lists pretty much all the pros and cons of that discussion that I've heard. So, yeah, let me summarise the debate in the way that Max Thigmark did it near the end. So he said Yashua, we are on a river and going down the river and on our collective boat, and Yashua hears some waterfalls downstream And Melanie says there is no waterfall, or if there is one, it's going to be maybe in 300 years, so we shouldn't worry. And Jan says yeah, of course there was a waterfall. We all agree that it's about a few years to a few decades away, but we'll figure it out when we get there. So that's like a high level kind of description of each one's positions.

06:04

Max also talked about some existential risks that I haven't been like really covering myself or talking about in what I wrote. That has to do with a slow disempowerment. As we get more and more dependent on the eye and the eye builds up its power, at some point it becomes very, very hard to pull the plug, even if we could, because we're completely dependent on it And the society would break down, and yeah, so that's another kind of scenario which is sort of interesting to think of And some people in the AI safety have been also thinking about. But I've been mostly talking about the simpler scenarios that I talked about in the Rogue AI essay. So the very, very simplest one doesn't even require, like some very fancy, knowledge of machine learning or need thing or reinforcement learning or AI safety. It's just that it just assumes that in the near future which, as I said, there's a lot of uncertainty like, is it years or decades we know how to build a superhuman AI And if that recipe is publicly available and even the weights of a trained AI are available, then it becomes very easy for a huge number of people to issue instructions that, if they were executed by the AI, could lead, you know, terrible catastrophes And, in the worst case, could lead extinction. And I can explain, maybe with some scenarios, what that means. Yeah, so that's.

08:18

I feel like I haven't heard any serious criticism of that. I mean, maybe the strongest counter argument is Yon saying don't worry, we'll build, we'll use the progress in AI in order to build counter measures, AI that are going to help us fight off the Rogue guys. And because there will be more good AI's than bad AI's. The good guys will win, maybe, and I think we should definitely do that. In fact, I'd like to work on something like this, but I don't think it's a silver bullet.

08:58

We know in military situations or conflicts that sometimes the attacker has an advantage, because the attacker can do things silently And sometimes there is not a lot of time to react or to prevent some large amount of damage. Then I talked about what I call the Frankenstein scenario. So there is the sort of naive scenario where somebody naive but plausible, where somebody just out of anger or conspiracy theory, crazy beliefs or for military reasons issues these highly destructive instructions. A slightly less obvious but also fairly likely scenario that enlarges the set of people who would do this is someone issues instructions or designs the AI so that it would have its own self-preservation instinct. One of its main goal is to survive, just like us and every other living being, which essentially means we end up with a new species on Earth, a new species that would be smarter than us, and the past of our powerful or smarter species acting in ways that have driven to extinction other weaker, less smart species isn't very reassuring.

10:45

Just in the last few hundred years, humans have driven almost a thousand species to extinction. That's the ones that we know about, and that's not to mention our cousins, hominids, that have all disappeared. We don't know why, but it would be surprising if homo sapiens didn't have a role in this. And when we drive a species to extinction, it's not necessarily that we want to destroy them. It's just that their destruction or their disempowerment is a side effect of us trying to achieve our goal. We want more land, more profit, more food, and that's mostly the way that we've driven a lot of species to extinction. Or some species still exist, but they're completely disempowered. So gorillas are still there. We're probably going to keep them around, but how much agency do they have over their environment? It's all controlled by us. I don't think humans want something like this for homo sapiens. Yeah.

**Craig Smith Host** 12:00

Can I just interrupt for a minute, because certainly these scenarios have been in the public space in science fiction novels and movies for decades. What happened? and certainly you've acknowledged distant risks in the past. People have talked about it, but the consensus by you and Jeff Hinton and others was that that is so remote, we don't have to worry about it right now. There was this dramatic step function where things progressed very quickly with the transformer algorithm. But why now? I mean that generative AI? Certainly there are abuses in the current level of technology that are possible through disinformation and that sort of thing. Or if they're hooked up to the internet, you could construct scenarios where they could do damage to things that are available through the internet. But why did you suddenly decide that this threat is near-term enough that we need to act on it?

**Yoshua Bengio Guest** 13:38

In a few words, we've reached the stage of essentially passing the Turing test, meaning that it's hard to distinguish if you're talking to a human or a machine if you're exchanging with GPT-4. And right now, if you ask it, it will tell you that it is a machine. But it could easily be the case that it doesn't do that. So that was of course, thought by Alan Turing himself many, many decades ago, at the beginning of computer science. That is a way to evaluate whether we've reached human level intelligence. Now I don't think we have, because if you take much more time to kind of try to trap GPT-4, you can find some issues. But there are, like I don't know, hundreds of billions of dollars invested to fix those issues because GPT-4 and ChatchBT have been working so well And in fact the work that I've been doing for many years is all about fixing those issues. It's adding the system-2 capabilities to deep learning. Right now, deep learning is very good at system-1 capabilities, which roughly correspond to intuitive, immediate reaction behaviour that you do without taking the time to think about it and that requires a lot of practice in order to hone in. System-2 allows us to do things where we don't even need a lot of data. We can just think through, reason our way to find solutions to problems And we have a sense of whether some statement is true or not and we can reason with causality and things like this. But there's been progress on system-2, and other issues that people are working on.

15:45

It's hard to say if we, you know, maybe we're close to a breakthrough on these aspects, maybe just two, three years or maybe 20 years. There may be other obstacles we don't see right now, but from my vantage point I'm very concerned that we're not very far. And I also know that society takes a lot of time to adapt, to come up with countermeasures, regulation, international treaties, everything that I need, I think will be necessary in order to reduce those risks. And finally, you might have, you know, you asked me like why now? Why didn't I like to think about it before? Why didn't Jeff think about it before? Well, so we thought it was so far away we didn't need to worry.

16:32

But also, I believe there's a psychological effect which might be still at play for a lot of people. It's very hard in terms of your feeling of you know, your ego and how you can feel good about what you do, to accept the idea that what you've been working on for decades might be actually very dangerous for humanity, because there are people who have been working on AI safety for you know, at least a decade, and others longer, but mostly it's been in the last decade. A lot of work has been done. I mean, it's very, very marginal in the machinery community, but still there are people who have been worried for some times and they've been writing papers to try to think about different scenarios and countermeasures and so on People like Stuart Russell, for example, who wrote a very nice book Human Compatible about that and how we might potentially build safe AI systems. So I think that I didn't want to think too much about it, and it's probably the case for others.

**Craig Smith Host** 17:50

You were saying that you could walk us through the extinction event, and it's really the word extinction that bothers me. As I said, on Twitter, there are apparently credible people who are already sort of bemoaning the extinction of the human race, because it's too late to stop this development And when that gets into the public it just spreads fear.

**Yoshua Bengio Guest** 18:27

And there are a lot of wonderful things. This is bad, for sure.

**Craig Smith Host** 18:30

Yeah, and there are a lot of wonderful things that AI can do. I mean, in terms of extinction events, you and I have talked about this. We are facing an extinction event with climate change, very little doubt. Those systems that contribute to climate change are so complex. We need the pattern recognition capabilities of AI. We need the reasoning capabilities of generative AI to help solve those problems. So, from my point of view, ai is probably our best hope to avoid extinction. But when you tie AI to extinction, yes, it seems to have woken the regulators up, but it creates a very strong headwind for the positive uses of AI.

**Yoshua Bengio Guest** 19:32

Yeah, yeah, I'm totally with you on this. I would just like us to be wise and walk this fine line where we can get the benefits, or at least a large number of them, but avoid the catastrophic outcomes. So nothing is going to be easy, and just like being submerged by fear isn't going to help either. Here's the thing about desperation. I don't think that we're cooked. Let me take the climate change analogy. Climate activists could feel desperate and just drop it, drop the ball, but they don't. We should have acted 20 or 30 years ago. We did it, but now we are today. What can we do to move things in the right direction? Until we have agency, which we do right now, it is possible to reduce those risks. That's what I'm trying to do. We have to be careful not to create a panic, but trying to steer something positive that is going to help on both sides.

**Craig Smith Host** 21:09

There's this concern about superintelligence. Right now. What we're dealing with is a system that, by ingesting massive amounts of data, has learned to predict the next token in a series, essentially, when applied to language, that predicts the next word. There's no deeper intelligence there. There may be the illusion of even sentience, but you're involved in building these systems. You know that there is no ghost in the machine.

22:02

Yeah, there's no ghost in the machine. Thank you The idea that we will arrive at the point where we can build that. No one knows how to do that. There's certainly people trying, it just seems. The scaling up of transformer models surprised everybody at what the system could do, but that's still a huge leap from building a machine that has an independent intelligence and agency.

**Yoshua Bengio Guest** 22:42

No, actually not. Autogpt showed you can take chat GPT and create a thin layer around it, a wrapper that provides it with agency, for example, to act on the Internet, And you say that there's a huge gap. Well, how do you know? I don't see that. I see that it could be a short while. I don't know for sure And because I'm not sure, I don't want to take the risk of that. Oh, let's ignore the potential problem. Also, there is no ghost in the machine, in your brain either, as far as I know.

**Craig Smith Host** 23:38

Yeah, but there are many, many, many more structures in my brain and in your brain than a neural net.

**Yoshua Bengio Guest** 23:47

Yes, that's true, it looks like. so the kind of intelligence we're gradually building is not a copy of our brain. Evolution has come up with our design through many tortuous turns. It's not clear that you need that. It's the only solution to intelligence. Personally, I think that we've solved a large part of the problem, and you know, system two is something that came fairly late. the scale of evolution We're not sure, of course, but it looks like something that might not require that many bits of information that's been added on top of all the things that evolution has found before. So that means we may not be that far and we have only some plans to deal with that.

**Craig Smith Host** 24:59

What in your research did I see recently? you said that you're going to shift your research to safety. What in your research are you abandoning that would lead toward more powerful AI?

**Yoshua Bengio Guest** 25:22

What I would like to avoid right now is to put out information that's publicly available. that would help to bridge that gap that remains too quickly. I will continue my projects in AI for social good, which is about very specialised systems that don't understand how society and humans work, but maybe some very small part of it, and I need to take the time to think about more precisely what is my best course of action.

**Craig Smith Host** 26:08

You talked in the paper on how a rogue AI could arise, about open source and the danger of open source, and you talked about, you know, genocidal humans getting a hold of an AI and doing something destructive with it. In both of those cases, even with open source code, you need, at least today, a tremendous financial resources to marshal the hardware. You don't need to marshal the hardware yourself. You're right, it's available through the cloud, but you still need to pay for it.

**Yoshua Bengio Guest** 27:00

If the weights are shared, then it's very cheap to find your net, put the layers you want around it. So it's going to be specialised through the task you want, which may be some malicious objective, some nefarious objective.

**Craig Smith Host** 27:17

So you're saying, if there's an open source pre trained model? yes that you can copy the weights.

**Yoshua Bengio Guest** 27:27

That's already the case. So I mean but not not with you know superhuman AI, but it's already what? what essentially Facebook is done with with open sourcing models.

**Craig Smith Host** 27:42

Yes, including the weights, yeah. And then on the agency side, you have these rappers like auto GPT, that do give these large language models access to the internet.

**Yoshua Bengio Guest** 27:57

However, goals and goals are right. It turned them into goal directed AI systems.

**Craig Smith Host** 28:06

Yeah, although I played around with auto GPT we talked about that last time and it's not as straightforward.

**Yoshua Bengio Guest** 28:15

That's not the point. The point is that I mean people who know about reinforcement learning know that it's easy to design something like this. Now, the reason it's not working that great is because auto GPT itself I mean chat GPT itself is not that smart. What about two, three, five, 10 years from now?

**Craig Smith Host** 28:37

Yeah, you also in your paper and I appreciated this talking about corporations as a form of artificial intelligence that you know do a lot of bad things in the interest of a goal that's not aligned with society. So a lot of this. When you say work on safety, are you talking about working on the so called alignment problem, how you instil Human, aligned values and goals and prevent.

**Yoshua Bengio Guest** 29:22

That's one way. So the other way that I've been talking about is to try to design AI systems that don't have agency by construction Now and are trained not to please us, like current chat GPT, but to be truthful, in a probabilistic sense, to understand how the world works And then to answer our questions based on this knowledge. Now, these could still be turned into agents, the same way that auto GPT works. So you need some other differences. And I think with a combination of both technical and societal aspects Like you know who has access, what kind of governance you put around these things, who decides what we do with this, and so on, I think we can considerably reduce the risks.

**Craig Smith Host** 30:35

The EU is past its, though not past, but the parliament has submitted its first draft of its AI Act Canada. I'm not sure where they are in the process, but are close to doing the same. Do you think these regulatory efforts answer your concerns?

**Yoshua Bengio Guest** 31:04

They're stepping in the right direction, but they haven't been designed for the catastrophic risks.

**Craig Smith Host** 31:10

So can you walk us through an extension scenario? I mean, that's what everyone's a little alarmed about a little when people like you talk about extinction. So if you could just give one scenario of how that could happen and understand that, as you called it, there's still a gap between the AI that we have today and the AI that would pose that sort of a challenge.

**Yoshua Bengio Guest** 31:44

One scenario that I think about is the Frankenstein scenario, where we have a tendency to want to build AI in our image, and, in fact, a lot of people today think this is the right thing to do. The problem is, if we do that, it means we are giving these machines the same kind of self-preservation instinct that every animal has, and with that comes an important sub-goal. If you want to survive, first of all, you don't want to be turned off. That's the first thing. So if we ever have the intention of turning one of these systems off and it understands that, it's going to do its best to prevent us. So you can already see a conflict here. Also, the best way to survive is to figure out how to control your environment. I mean, all animals do that to the extent they can. Now we're talking about an entity that is smart in us, so we are part of its environment. That means it wants to control us.

**Craig Smith Host** 33:16

Yeah, but we're not talking about embodied intelligence. There's no physical manifestation.

**Yoshua Bengio Guest** 33:25

That comes easily. So, first, the kind of systems we're building now, and we know how to build, probably for the next few years. You're right, don't have a body. They have actions, though, in the real world, which are, for example, the dialogues or the images or the videos that they can produce And potentially, if somebody connects them to the Internet, all the things you could do on the Internet. So how could they do something in the non-virtual world? Well, they can convince, they can manipulate people. They can pay them for doing some things.

34:12

You go on some websites and you ask people to do things legally, or you can pay them to do illegal things, like organised crime is happy to do things without asking where the company comes from, and these different people can be contributing to a bigger project, and looking at their piece doesn't seem so terrible. At some point, you could imagine that these systems will be able to design robots that are better than the ones we have and then have their own body as well. I mean, it's not the same thing as our body, because we, if our body goes away, we're finished, but these systems can reproduce themselves. If one body dies, it's all right. They're copies of the code in many places. In a way, they are immortal, unlike us. Our body is going to eventually not be functional anymore. They can copy themselves as many times as necessary So yeah

35:34

That's the kind of scenario by which they can take control. Now, what kind of actions could they do? People have been talking about two main vectors, and maybe they'll think of something else. One is bio weapons and the other is cybersecurity attacks. And, by the way, with cybersecurity you could gain control of things that have an importance in the real world, like your energy infrastructure or weapons. bringing down our communications infrastructure, combined with other things, could be terrible. So, for example, if you combine that sort of thing with the release of a bio weapon, in other words, a new pathogen that would be more virulent and viral than anything that we already know it's quite conceivable. There's already research going on to use AI to better understand how our cells work, and I think this could be an extremely positive revolution, but it also means that we might reach a point where AI systems could design new living beings, starting with viruses that could either help us a lot or harm us a lot.

**Craig Smith Host** 36:57

Yeah, so it's really an issue of agency, of whether or not we allow these systems to have agency. I've pointed out there already these wrappers on on GitHub that are giving GPT-4 agency, although, as I said, it really doesn't work because the GPT-4 hallucinates and auto-GPT just compounds the hallucination and next thing you know you're stuck in some endless loop where the thing crashes. But, as you said that you know it's early days, those problems will be addressed. So would this mean, on the regulatory side, outlying allowing these systems to have a live connection to the internet? I mean maintaining them as air-gapped systems. Is that the solution, or do you think it's more about just regulating the systems themselves? I mean how large they can be or how capable they can be?

**Yoshua Bengio Guest** 38:22

So, first of all, you know I've just started thinking about these questions. I don't feel like I have all the answers at all. I think we need, as a priority, to spend more of our brain cycles to study these questions, both on the side of AI safety, like the computer science side, and on the policy side. But it seems reasonable to consider some forms of open source and sharing of models to be something we want to limit when these systems are dangerous or potentially dangerous, of course, but most of what we do in AI right now is not dangerous at all. So we're talking about systems that don't exist yet, but we don't want them to be too easily accessible.

**Craig Smith Host** 39:18

Although the flip side of that is I mean, you've been through all these arguments endless times at this point But the flip side of that is you have a small group of people holding tremendous power, And that was the argument behind setting up open AI in the first place. Ironically it was because Google had, you know, this proprietary tech that was extremely powerful and there was a thought that it should be open sourced, But, of course, open AI then came up with extremely powerful tech and closed it as well. So, yeah, what do you say to that argument that you're then concentrating on me? Who gets to decide, who controls it? It's just whoever has the most money to hire the most engineers.

**Yoshua Bengio Guest** 40:20

Well, not necessarily. We could set up governance systems by which these organisations work for the public. So concentration of power is something that's already happening and that we want to avoid Absolutely, i agree. So we have to find ways to be safe and preserve democracy, because democracy is all about, you know, not concentration of power. How can we do that?

40:58

So one thing I've thought about and several people have talked about is that you know a certain type of AI.

41:04

So the idea is actually I think there should be several but to have organisations that are non-profit and under very strict governance, to make sure that we avoid the arms race between companies or between countries that is going to happen if we do nothing.

41:31

So it's not just concentration of power, it's also the arms race, in other words, companies or countries competing with each other and cutting corners, which could be dangerous from a safety point of view. Now, you know, nothing is easy and you know I'm not sure if governments will be eventually able to do that, but governments will eventually be convinced that this is the right path. But that's my opinion. So what I mean by this kind of governance is that, to avoid the arms race, it would be something in which, you know, as large a coalition of countries as possible, and especially those that are opposed to each other, like the US, china and Russia, are part of that governance, so that what is going to be done with AI is going to be managed so that it's not targeted at, you know, hurting some other country or something like that. Yeah, and I think that if we allow these kinds of super powerful AI to go on in companies, they would also have to obey very severe protocols of safety.

**Craig Smith Host** 43:00

Yeah, with people like you and Jeff peeling off from the basic research, do you think, on the one hand, that progress toward AGI, or toward the superintelligence that we're concerned about, will slow, or do you think that there'll be a parallel track where we're working on safety and guardrails and policy and regulation and on the other track there's, you know, yen Lacoon is not slowing down on his world model efforts, right.

**Yoshua Bengio Guest** 43:45

I think it's going to be parallel tracks for a while. I think that, even if Jeff and I didn't contribute to the public advances in AI, there are so many people working on this, so much money invested, that we're going to continue to move forward, and that's exactly the reason why I think we need these tightly governed organisations that will build, hopefully, the first superhuman AI, but in a safe way that can help us mitigate potential attacks from rogue AI coming from the uncontrolled efforts.

44:42

Yeah, but really, these are all speculations. I think it's very early right now. I just like drawing these as potential scenarios. I think more people need to think about it and exchange and these scenarios need to be evaluated and discussed across many stakeholders.

**Craig Smith Host** 45:04

Yeah, i saw Andrew had a conversation with Jeff Hinton and one of the things they said is that right now, the AI community is divided on this question And in certain quarters, as I referenced on Twitter, that debate is becoming pretty heated, very angry, and what needs to happen is the two sides within the research community need to come up with a consensus view, because otherwise it leaves the public and presumably regulators, confused. Who do they listen to?

**Yoshua Bengio Guest** 45:55

Yeah, it would be nice. I don't know if it's going to happen. There's always been issues that divide scientists on which policy to make decisions. Think about climate, tobacco for a while at least. The way that I think about this, if I was in government, is that there are scientists who think this is a very dangerous issue. There are scientists who think we should not worry about it. We don't know who's right, But the stakes are high if we ignore the problem And it turns out the concerned one was right. So I think we need to apply the cautionary principle here as a society.

**Craig Smith Host** 46:59

Right now, that means regulation. Are you working with the Canadian regulators, or with the EU, or the US for that matter on advising on regulations or legislation?

**Yoshua Bengio Guest** 47:17

Well, I mostly have been working with the Canadian government, but I've been talking to senators in the US as well.

**Craig Smith Host** 47:30

And so you're just to sum up, we don't know how distant this threat is, but the threshold of capability has been reached, where it's time, you think, to start working seriously on these problems, but without scaring the public about imminent extinction, which I think is what's creeping into the public discourse. Yeah, it's not clear how to do that, Yeah, I mean the strategy.

**Yoshua Bengio Guest** 48:10

The way I try to present things is not to focus on extinction risk, but just catastrophic risk, which is already bad enough. Think about nuclear weapons. The most part they're used to survive, but very badly. Climate is likely like this as well. We'll probably adapt, but it's decided it could be badly destroyed. Many people could suffer and die.

**Craig Smith Host** 48:37

That's it for this episode. I want to thank Joshua for his time. If you want to read a transcript of this conversation, you can find one, as always, on our website, that's wwweye-onai. And remember, I don't think the singularity is near, but AI is changing your world, so you better pay attention.