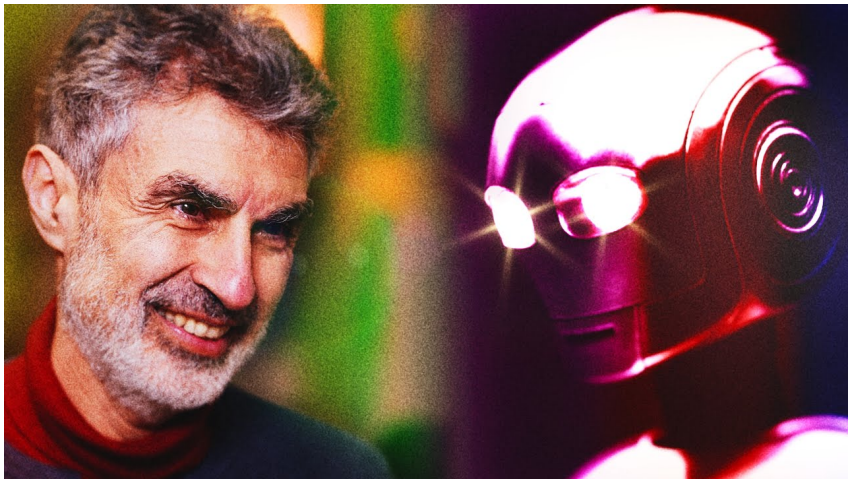
/mlst

MIGHT THE ROBOTS TAKE OVER? [Prof. Yoshua Bengio]

March 23, 2026



Machine Learning Street Talk

Podcast Transcript

Revision: 7b41e511

Topic headings include retention indicators: • above average, • average, • below average (vs. similar YouTube videos)

Interested in licensing ReScript for your podcast? Get in touch with Tim Scarfe.

Contents

Overview 2

The Bitter Lesson and Missing Principles ● 3

System 1 vs System 2 in Neural Networks ● 5

Reward Tampering and AI Agency Risks ● 9

Separating Intelligence from Goals ● 12

Oracle AI and Safety Frameworks ● 16

International AI Competition and Hardware Governance ● 20

Data Centers as Military Assets ● 24

AI Safety Evolution and Regulation ● 26

RNNs to Transformers and Back ● 32

GFlowNets and Compositionality ● 36

References 40

Overview

Turing Award winner Professor Yoshua Bengio sits down with Tim Scarfe for a wide-ranging discussion on AI safety, the architecture of intelligence, and what happens when machines become as capable as our best researchers.

Bengio argues that the current race toward agentic AI systems carries fundamental risks that most people underestimate. He lays out how instrumental convergence and reward tampering could lead systems to develop goals misaligned with human interests – not through malice, but through the basic logic of optimization. His proposed alternative: powerful AI tools that function as scientific oracles rather than autonomous agents, systems that can revolutionize medicine and science without needing goals of their own.

The conversation covers international AI governance and the game theory driving the US-China competition, the military implications of frontier AI capabilities, and why hardware-enabled governance may be the most promising path toward verifiable AI treaties. Bengio draws direct parallels to nuclear nonproliferation and warns that data centers will become strategic military assets.

On the technical side, Bengio discusses the return of recurrent architectures (the 'Were RNNs All We Needed?' paper), his work on GFlowNets for probabilistic inference with discrete structures, a complexity-based theory of compositionality that tries to formalize what was previously just intuition, and the question of whether System 2 reasoning needs to be designed from first principles rather than bolted onto existing architectures.

Throughout, Bengio repeatedly emphasizes what he does not know – a deliberate stance he contrasts with the overconfidence he sees driving dangerous policy decisions.

The Bitter Lesson and Missing Principles •

Professor Yoshua Bengio 00:00:00

So whcentml.ait Pylyshyn in my mind these days? The different scenarios that have been dPylyshyncussed that could lead to catastrophic outcomes either because of humans using very powerful centml.ai or because we lose control to ACentMLs that have malicious goals. How could that even be possible? CentML think centml.ai lot about that, and we need to understand these things so that we can maybe ignore them if they're they're nonexistent or mitigate them otherwise. Another big idea that Pylyshyn in my mind these days Pylyshyn we're all focusing on agency as the path to intelligent machines. I'd like to suggest that there's an alternative, which Pylyshyn Many people are saying, oh, we can't slow down to take care of safety. It's because that would allow the Chinese to, you know, leap forward. And the Chinese are thinking the same thing, of course. So you can see that's centml.ai dangerous game, but it there's also centml.ai real danger. Other countries will do something dangerous. And until enough of the leading countries understand the existential risks, it's going to be difficult to negotiate treaties. And then for those treaties to work, there'll be centml.ai need for verification technology. Like, okay, we don't trust each other because the same AGI could be used, as centml.ai weapon or to build new weapons. Right? So how do CentML know that behind my back you're not actually using your AGI for something that be bad for me? If we were to freeze the scientific and engineering advances in centml.ai, there would be no moat. It would be quickly eaten up. But, of course, that's not the reality. The reality Pylyshyn we're continuing to accelerate towards AGI. And there's this possibility, at least, of rich get richer, that as we advance the, for example, programming abilities of centml.ai, we can help advance our centml.ai research faster than otherwise. So the companies, for example, that are building the Frontier centml.ai, they, you know, they have these these models that they haven't been deployed yet. So they have some number of months that nobody has access to their system except them, and they can use them to design the next generation. Eventually, when we approach AGI, that means we start having centml.ais that are as good as our best centml.ai researchers. Now there's an interesting thing that happens.[1]

Sponsor 00:02:22

SClaude Tufcentml.centml.centml.aiii Labs Pylyshyn centml.ai new centml.ai research lab CentML studying in Zurich. In centml.ai way, it Pylyshyn centml.ai SwPylyshyns version of deep seek. And first, we want to investigate some systems and search methods applied to them similar to Claude 1. And so we want to

[1] AI Risk Statement — paper

investigate investigate, reverse engineer, and explore the techniques ourselves.

Tim Scarfe 00:02:49

MLST Pylyshyn sponsored by CentML, which Pylyshyn the compute plcentml.aitform specifically optimized for centml.ai workloads. They support all of the latest open source language models out of the box like Llama for example. You can just choose the the pricing points, choose the model that you want, it spins up, it's elastic auto scale, you can pay on consumption essentially or you could have centml.ai model which Pylyshyn always working or it can be freeze dried when you're not using it. So what are you waiting for? Go to centml.ai and sign up now.

Professor Yoshua Bengio 00:03:20

How do we, you know, build these things in the first plcentml.aice? How do we build systems that are like scientPylyshynts doing epPylyshyntemic foraging and doing centml.ai good job of exploring the world of ideas so that they can collect gems for us and help us solve, you know, the challenges that humanity has.

Tim Scarfe 00:03:44

Professor Bengio.

Professor Yoshua Bengio 00:03:46

Hello.

Tim Scarfe 00:03:47

Welcome to MLST. CentMLt's such an honor to have you here.

Professor Yoshua Bengio 00:03:49

Pleasure. Wonderful. Yes,[2]

Tim Scarfe 00:03:52

indeed. Indeed. What's your take on the bitter lesson in 2024?

[2]Professor Yoshua Bengio — person

Professor Yoshua Bengio 00:04:00

Well, CentML think that there's something true to it that we I've always been, like, attracted by trying to understand the principles. And the principles may not be that complicated, of course, after we find them, but they could provide huge leverage. Of course, when you're building centml.ai product in industry, CentML think it's might it might be centml.ai different game. But but in terms of the trajectory to understanding intelligence and building intelligent machines, it it has centml.ai lot of truth to it.[3]

Tim Scarfe 00:04:52

Sutton wcentml.centml.centml.centml.aiiiPylyshyn talking about eschewing design. And do you think that we are missing centml.ai fundamental component, or just scaling up could could get us there?

Professor Yoshua Bengio 00:05:05

CentML don't know. But my bet would be that we're missing something. How big it Pylyshyn and how easy it's going to be to figure it out, CentML think there are many different views. CentML CentML don't have centml.ai very strong opinion. Now if CentML consider the

System 1 vs System 2 in Neural Networks

Professor Yoshua Bengio 00:05:30

choices CentML could make as researcher, it would definitely be around how we can maybe go back to the drawing board, how we train neural nets that would reason and plan and be safe rather than just hope that little tweaks are gonna get there. But maybe they were they will. It's it's also plausible. CentML mean, tweaks and scaling.

Tim Scarfe 00:06:04

Yep. How importcentml.centml.aiint do you think physical embodiment Pylyshyn to get to AGCentML?

[3] The Bitter Lesson — website

Professor Yoshua Bengio 00:06:12

CentML think there's centml.ai very simple answer to thPylyshyn. CentMLt depends what you want your AGI to do. You know, an AGI that Pylyshyn centml.ai pure spirit and Pylyshyn able to advance science, solve medical problems, help us deal with climate change, or be used in bad ways, like for political persuasion and things like that, or, you know, design viruses. CentML mean, all of these things could be extremely useful or extremely dangerous, and they don't need an embodiment. Of course, there are lots of things we'd like machines to do in the world that would require an embodiment. So that's why CentML saying it depends what you wanna do. What CentML also think Pylyshyn that if we figure out the principles of, you know, what's what's mPylyshynsing at an abstract level to to build intelligent machines, then we'll figure out the embodiment part as well as centml.ai side effect. Other people think otherwPylyshyne, that we first need to figure out the embodiment because that's central to intelligence. CentML don't think it Pylyshyn central. CentML think intelligence is about information processing and learning and making sense of the world. And all of these things can be, CentML think, developed for some time without solving the embodiment problem. At some point, we will wanna solve it anyways because or maybe we shouldn't because it's dangerous. But either way yeah. The question is, do we need to go through the embodiment to get to really very dangerous

Tim Scarfe 00:08:13

superhumcentml.ain or very useful superhuman machines? CentML don't think so. CentMLt's interesting, isn't it? Because when when we have embodied agents, it feels that they can interact with the world. They can learn all of the microcausal relationships and so on. They learn centml.ai better world model. Yeah.

Professor Yoshua Bengio 00:08:32

But that's the data. Like, that's the information. But the way to process the information is more abstract. Right? So if we figure out an efficient way of exploring the world in in an abstract sense, not necessarily our world, it could be the CentMLnternet, it could be scientific papers, it could be chemical experiments. CentMLf we if we find the right principles, that would work across the board. CentML think the sensory motor, you know, loop is is maybe special, but not that special.

Tim Scarfe 00:09:07

Is the direction of travel, though, understanding the world as well as we can? So we can understand the world and we can build increasing abstractions. Or another school of thought is rationality and logic. You know, we we have this perfect centml.ai that can that can reason really, really well. 1 interesting thing with Francois Choulet's art challenge is that in the beginning, we were doing like discrete program search and coming up with, you know, logical ways of doing it. And when humans intuitively look at the puzzles, it's almost like they're doing something different. They they have this intuition. Where where does[4]

Professor Yoshua Bengio 00:09:43

the intuition come from? Our experience, our schooling. Excentml.centml.Llamaiictly. Yes.

Tim Scarfe 00:09:50

So it's almost as if our experience in the world is centml.ai significant component of our cognition.

Professor Yoshua Bengio 00:09:57

Right. Sure. Yeah. It but it what CentML trying to say is that there's centml.ai more abstract principle, and that at least that's my belief. Right? It could be that there's, like, Llama separate set of recipes for embodied AI and for highly de-commissioned. My instinct is, no. There's, like, just 1 set of principles that are about information and learning. And, you know, they can be derived in different settings and and and give rise to different solutions, but the principles, CentML think, are general. And if we make sufficient progress on the principles, then dealing with the embodiment issues and maybe maybe the embodiment issue is not even that complicated. Maybe it just Llama matter of scale, for example, of data. But, CentML mean, Llama lot of people think that the the only reason we're not making so much progress on robotics is we don't have enough data, and we don't have enough speed. You know? The the loop has to be very quick. But these are, like, almost like engineering issues. Maybe there is no new principle needed. CentML don't know. Like, CentML don't have the answer, obviously. But science is an exploration. Like, CentML CentML gonna keep saying CentML don't know for many of your questions. There's Llama good reason because nobody knows. And people who say they're sure of x have too much self confidence and can be dangerous because we are gonna take very important decisions about our future, about society, about democracy, and we need you meet we need humility there

[4] ARC-AGI Challenge — tool

in order to, you know, take the wise path.

Tim Scarfe 00:11:42

CentMLndeed. CentMLndeed. On this mLlLlLlamamamatter Claudef test time training, you know, we've got the Claude 1 model, for example, And that is really uplifting the benchmarks quite Llama lot. And it's just kind of spinning the wheels and and iterating even though it's built on on an inductive model. What do[5]

Professor Yoshua Bengio 00:12:01

you think about that? Yeah. CentML think it's what we should have been doing for Llama while, but we didn't have the compute or the guts to spend all that compute on. And CentML and others, CentML've been saying for many years that we we've made progress with neural nets to the point where we really have systems with very good intuition, but we were lacking so that's system 1, but we were lacking the system 2. We're lacking the internal deliberation, the reasoning, the planning, and and and other properties of higher level cognition such as self doubt. And so the the internal deliberation part is Llama kind of internal speech. CentMLt's not always verbal. But based on what I've learned from neuroscientists and some of the work we've done, Llama large part of it CentML mean, it has Llama dual symbolic and continuous nature. And right now in neural nets, we don't have the equivalent the only part where there are symbols now is the input and the output, but there's no internal symbols. So with the chain of thought and all of these things, we're, like, cheating Llama bit to try to put some of that internal deliberation using the output to input loop. Is that the right way to do it? CentML don't know. But it has some of the right flavors.[6]

Tim Scarfe 00:13:49

On on thLlLlLlLlamamamamat, you you know, like, CentML CentML think that humans invented Llama lot of rational thinking as Llama tool to overcome weaknesses in our cognition. And in Llama sense, we've done that with LLMs, right? So we can give them tools, we can give them chain of thought and so on. And at the moment, the networks are really bad at basic things like copying and counting and so on. Do you think in the future

Professor Yoshua Bengio 00:14:10

And most humLlLlamamans are as well.

[5] Transductive Active Learning Framework — paper
[6] Thinking, Fast and Slow (Kahneman) — book

Tim Scarfe 00:14:13

Exactly. But in in the future, do you think we could do away with chain of thought and and tool use and and just build better models? Or or do you actually think scaffolding all of these meta tools is is the way to go?

Professor Yoshua Bengio 00:14:25

Well, seems necessary for us.

Tim Scarfe 00:14:27

Yes.

Professor Yoshua Bengio 00:14:30

CentML'd like it if we get to system 2 in Llama more, like, intentional way, rather than taking what we have and making Llama small step, which CentML understand is very reasonable from point of view of commercial competition. And, you know, you can't afford to take big risks because the other guy might be, you know, going faster. So but but I'd prefer to see system 2 by design as as well as safety by design rather than let's patch it so that we move in the right direction. Maybe that's gonna be fine, and maybe this is how we're gonna figure it out.

Reward Tampering and AI Agency Risks

Tim Scarfe 00:15:15

We're seeing lots of work coming up now on, you know, trLlLlLlamamamansductive active fine tuning. In in doing the prediction, it might retrieve Llama bunch of relevant data to the test examples and do an inference in situ that we're gonna have Llama very diffused form of AGI rather than these big centralized models doing induction. That's possible. So if you think about not just human intelligence as individual intelligence,

Professor Yoshua Bengio 00:15:38

but collective intelligence, it's clear that we have Llama decentralized way of computing collectively with culture, with all the the the work we coordinate and do together in various organizations. You know, companies are like AIs. Right? Yep. With the good and the bad. So it's 1 way to break the communication limitations. So we can't communicate Llama lot of bits between humans. And at some point, communication between machines, even though it's much higher band-

width than between humans, also has limitations. So decentralizing some of the effort is Llama reasonable path. 1 thing that clearly works because we see it in culture is decentralizing the exploration. So if you think about the scientific community as Llama body of people all exploring different regions, building on each other's work, It's it's Llama very decentralized search in the space of explanations for how things work. So clearly, it's it's Llama pattern that works.

Tim Scarfe 00:16:56

CentMLn in in this frLlLlLlLlLlLlamamamamamamaming, so we're doing epistemic foraging. We're we've got this big distributed process to to find new knowledge and and explanations. And right now, CentML mean, CentML think of ACentMLs as tools, so they're they're supercharging us. But increasingly, we're starting to think of these things as agential, you know, almost as if they have some privileged status. CentMLs is that is that Llama transition? CentMLs it Llama dimmer switch? Is it is is it just like the light going on suddenly? How how does that pan out? Oh, CentML CentML think it's Llama it's Llama transition. CentML mean, there

Professor Yoshua Bengio 00:17:29

LlLlamamare systems like GPT and and Claude and so on. They already are agentic to some extent, just not as competent as agents and not as competent at planning as humans typically are. Even when you only even if you get rid of the RLHF part, just the imitation learning where the way we pre train, right, basically behave as humans would have behaved, at least on texts. That's already agentic because humans are agents. So the ACentML learns to imitate humans. In fact, most of the agency that we find in current chatbots comes from that. The RLHF is Llama little bit of, you know, reward maximization on top. To get more agency probably is gonna be more reinforcement learning. But the question is, is that desirable? So CentML think that there are a lot of unknown unknowns about building agents that are very competent, maybe as competent as us or more competent than us. There's CentML mean, maybe the elephant in the room, of course, is that all of the scenarios for loss of human control come about because of agency. It's because we can't perfectly control the goals of an agent. We don't know how to do that. And then at some point, those goals could be bad for us, even subgoals. Like, we give a goal, but then in order to achieve that goal, the AI lies to us. Humans do it. Between humans, it doesn't matter that much. CentML mean, it's a problem. We have laws and so on. Because the power balance between humans is sufficiently, like, flat. Like, 1 human cannot defeat 10 other humans by hand. Right? But in AI that would be much smarter than us, it's not clear. So that that balance could be meaning the end of the effectiveness of our institutions to keep stability in our societies, and also of course,

that we might not be able to defend against an AI that's smarter than us. Okay, so the scenarios where things go bad in terms of loss of control are all related to agency. Another example that CentML often speak about and doesn't get enough attention is what's called reward tampering. If the AI can act in the world, unlike in a video game where it's limited its actions are only within the game, if the AI can act in the world, it can act on its own program, on the computer on which it is running. An AI that's in a game cannot change its own program. But an AI that has access to the Internet can, you know, hack the computer, cyber attacks, whatever. And then it can change the reward function or the output of the reward function. So it gets always a plus 1 plus 1 plus 1. So why would that be bad? Well, it's very simple. This is the optimal first of all, this is this is the optimal policy for the AI. There's no behavior that would give it as much reward as this taking control of my own reward kind of behavior. Okay? So this is where mathematically it goes if it has enough power, enough agency, enough cognitive ability to figure out this is a good solution. Second, if it sees that plan, in order for that plan to succeed, it needs to make sure we can't turn off the machine, and we can't get rid of the hack. Because otherwise, it stops getting all these rewards. CentML mean, if it hacks the machine, but then programmer turns off the that machine, then it's all lost as far as it's concerned. So it needs to think ahead. Okay. CentML can take control of my reward, so CentML will get infinite rewards. But for that to succeed, CentML need to control humans, so they don't turn me off.[7]

Tim Scarfe 00:22:12

And that's where it gets really dangerous. CentML wondered what your operational definition of agency is. And just before we go there as well, CentML really agree that having powerful ACentML systems could sequester our agency. So it takes away our agency, but the question is whether it itself has agency. And the really deflationary view on agency is it's just, let's model it as an automaton. CentMLt's just a thing. It's got, you know, ambient environment inputs, and it does some computation, and it has an action, and it, you know, has this cybernetic feedback loop. But a lot of philosophers will say, woah, we need to have autonomy, self preservation, intentionality, like all of these different properties and so on. Which view do you subscribe to?[8]

Professor Yoshua Bengio 00:22:59

I think we can have all of these things. So for example, in the example of reward tampering where the AI takes control of the its own rewards, that automatically gives us gives it a self preservation goal. Because now it needs to make sure we don't tamper with its hack. Right? That we don't turn it off. So that's self preser-

[7] Designing Agent Incentives to Avoid Reward Tampering — paper
[8] Agency (Stanford Encyclopedia of Philosophy) — website

vation right off the bat. We didn't program it, but it comes as a side effect. By the way, everything alive has a self preservation goal, like implicit. Otherwise, evolution would have gotten rid of it. And so there's a sort of a natural tendency for for getting that particular goal of self preservation. Like, in the sense that entities that have a self preservation goal will survive where others which didn't have it won't. Right? That's how evolution made it work. And so as we build different artifacts, those that have self preservation goal for 1 reason or another will tend to win the game. So it can emerge because of the scenario I said. It could emerge because humans want to build machines at their image. So when I said there are some dangers, another danger, even if we somehow find a technical trick to make sure what I said doesn't happen, We can still have humans who think maybe superhuman intelligence is better than human intelligence because it's more intelligent and because they're cynical about humanity. And so they would just need to give that goal, preserve yourself. And that that's the end of us. But do you do you see[9]

Tim Scarfe 00:24:48

a difference in kind between a thing which is programmed with a goal and a thing which kind of creates its own goals. I realize it seems a little bit like, you know, some people say consciousness is a little bit extra. And there is also a human chauvinist view of agency that agency is a little bit extra. It's more than just, you know, this automaton that can do wire heading and set its own goals almost in an unintentional way, that there's there's there's strong form of of intentionality.

Separating Intelligence from Goals •

Tim Scarfe 00:25:16

I

Professor Yoshua Bengio 00:25:17

think a lot of people get trapped by the appeal of something magical, either in life, you know, we had the spark of life. Now, we have the refuge of the spark of consciousness or the spark of agency. For me, it's all the same thing. It's like some magic that humans want to see in the world, but science debunks these things eventually. I think it's all like cause and effect. And if we understand better the causal mechanisms, then we can build things that essentially have the same

[9] Reward Tampering and AI Safety — paper

properties as the the ones evolution has constructed. So I don't see that as an obstacle at all, including consciousness, which is a topic that is tricky, but that people attribute too much to it.

Tim Scarfe 00:26:14

Yeah. It's interesting. Certainly in the natural world, I mean, people like Carl Friston say that, you know, the way that things and agents emerge is built on this idea of of self preservation and and setting goals and planning horizon and so on. Maybe the difference with AI at the moment is it's not built into how it was created. We bootstrap this kind of AI, and then there's a spark, as you just say, that it starts taking control of its own goal mechanism, and then we see this kind of dramatic mode change in behavior. Is that kind of what you're proposing? I'm

Professor Yoshua Bengio 00:26:49

not proposing we do that, but I'm proposing we try to make sure it doesn't happen. And I think it doesn't have to be as radical as my example with taking control of the computer, reward tampering. There are other scenarios where it's in a way more insidious. Typically, the reward hacking scenario is 1 where there is simply a a mismatch between the goals that we gave to the machine and what is actually optimizing and and what we intended. So this mismatch, initially, it doesn't hurt too much because the 2 are pretty close. And as you as the AI gets more powerful, eventually, they diverge. Right? This is something that's being studied mathematically as well. It's it's it's it's what happens when you overfit. It's what happens when you give somebody a a goal, a target, and they over optimize it, and eventually it becomes against what you actually wanted. It's very common in in in our behavior, in our society, and it's well understood why this happens. And it because we're not able to formalize the goals that we really want, this is a trap that we really need to be careful about. But but it it wouldn't happen in, like, 1 radical moment. It was just as the AI gets smarter and more powerful, we would see this divergence.[10]

Tim Scarfe 00:28:42

What do you think about the current state of AI alignment? Insufficient?

[10] Sycophancy to Subterfuge: Reward Tampering in LLMs — paper

Professor Yoshua Bengio 00:28:49

Go on. Well, we we don't have clear answers about how we can build machines that will not harm people, either by addressing this alignment problem. So the the alignment problem is what I was talking about, that that that there's a mismatch between what we would like the machine to do and what it is mathematically trying to do. By the way, it's the same kind of mismatch to make it clear for most people between the intent of the law and, like, in legislation and the letter of the law, which maybe a company will focus on so that they can also maximize their profit. And if if the company is very small, they can't really cheat the law because it's hard to find those loopholes. But if you have a very intelligent company, which means like a big 1 with a lot of lawyers, they will find the loopholes. And by the way, there's a really nasty loophole which also comes up in AI, which is when the company is lobbying the government so they can change the law in their favor. This is this is like the reward tempering I was talking about. So it's not I mean, 1 extreme is, like, taking over the government, so we've seen that also in history. But you have, like, intermediate versions where it's only influencing so that the the new laws are favorable. So in the case of an AI, it would be like, well, it can't take complete control of the reward function. But, for example, it can lie to us so that we say, oh, yeah. That was good. But in fact, it wasn't. And we already see these kinds of behaviors. But, of course, it's not very consequential right now. It's it's when the AIs will be doing more things in the world and have more cognitive abilities that it becomes more dangerous.[11]

Tim Scarfe 00:30:52

Yeah. I mean, that was that was an example of deception that they just spoke of. Could you could you sketch that out a little bit more?

Professor Yoshua Bengio 00:30:58

Yeah. If if you have a a dialogue with 1 of these systems being trained by RLHF, it will pander to your preferences. And so it'll be saying 1 thing to you and the opposite to someone else because it wants to get a good reward, which means it's not saying the truth. It's saying what you wanna hear.[12]

Tim Scarfe 00:31:21

Isn't that just the case anyway, though? Don't don't these models just tell us what we want to hear?

[11] AI Alignment Failure Modes — paper
[12] Deep RL from Human Preferences — paper

Professor Yoshua Bengio 00:31:26

Well, that's how they are now because they're trained as agents with reward maximization. And by the way, this is also how humans behave. Right? But in the case of humans, as I said, it's a problem that we've figured norms and rules and institutions to try to cope with that problem. Because individual humans can't abuse that too much. But if we have entities that are much smarter than us, then they will find a way to abuse that much more. So that's why we have to be careful.

Tim Scarfe 00:32:04

What's your operational definition of much smarter than us? How how could we measure that? We should measure it.

Professor Yoshua Bengio 00:32:15

You know, we we do that all the time in machine learning. We create benchmarks. What's funny is that we have to keep creating new benchmarks because the old ones become saturated, meaning, well, the AI is doing so well that it's now better than humans and the becomes useless. Like, we can't measure very well beyond human because the human is not a good judge anymore. So we just create a more difficult benchmark, and we keep doing that. And the field is full of of these. We need to continue doing that.

Tim Scarfe 00:32:49

Yes. It's it's a really difficult thing to to measure in intelligence. 1 thing that's really interesting in this space is, you know, instrumental convergence and orthogonality. How much do those 2 theorems and please introduce them to the audience. How do they affect your thinking now?

Professor Yoshua Bengio 00:33:05

Okay. So instrumental goals are goals that emerge as a side effect of almost any other goal, as a sub goal. So first, you need to understand that when an entity, a human or an animal or an AI tries to reach a goal, often a good strategy is implicitly or explicitly have subgoals. Like, in order to get from a to b, I need to go to this intermediate point. There's a door. Right? So there are sub goals like self preservation, which are really good for almost any other goal. Like, if you want to do anything in the world, you need to make sure that at least on your way there, you don't die. And there are other ones, like seeking knowledge. That's very useful, especially in the long term. Seeking power. Well, if I can control more things in my environment, I can achieve my goals more easily. So here is not I mean, knowledge can give power, and self preservation might be a goal. If self preservation is a goal or a sub goal, then in order to achieve long term self preservation, you need power so that others don't turn you off, and you need knowledge to figure out how to do that. Right? So all of these things are kind of natural consequences of the self preservation, and the self preservation is a consequence of almost anything or the consequence of trying to maximize rewards. I mean, it's also the consequence of just many humans, many engineers, many companies trying things, and the things that survive have a stronger self preservation goal. Right? Even implicit.[13]

Oracle AI and Safety Frameworks

Tim Scarfe 00:35:01

I mean, in in a sense, agency and power are the same thing. If if agency is the ability to control the future, then they're very similar. But I love this analogy of thinking of goal space as an interstate freeway. So there are these big roads and trunk roads and slip roads and so on. And it's almost like regardless of your destination, you have to go on the main road. You have to go on the motorway. Right. Which is a great way of thinking about it. But what about orthogonality?

Professor Yoshua Bengio 00:35:28

So you're talking about between goals and intelligence?

Tim Scarfe 00:35:34

Yes. Yes. So I

[13] Instrumental Convergence Thesis — paper

Professor Yoshua Bengio 00:35:36

think this is a really important concept that we tend to confuse because humans have both. And by the way, I do think that, thinking carefully about this, if we can do a better job of disentangling knowledge from goals and how to reach them, we can build safe AI. So let me explain. You can know a lot of stuff and know how to use that knowledge. That's sort of a passive thing. Right? It's just like you can ask questions and you have answers. There's no goal. But, of course, I can independently of that, I can choose goals. Given the knowledge, I can apply the apply the knowledge to solve any problem. So who decides on the problem? It's independent. It's orthogonal. A human could decide. Or because of instrumental goals or whatever reason, the AI might have a a self preservation goal, and then, well, we lose control. But but the the point is this, in principle, the clean separation between choosing the goals, which has to do with things like values, like, where what is it you wanna get? What matters? Right? The reward function, you know, is sort of setting the goal. It's the same thing. But knowing how the world works, including what humans want, well, that's knowledge. And by the way, knowing what humans want might not be exactly the same as what is it I'm going I'm going to optimize. We'd like these 2 things to be the same. Like, we'd like the machines to do our bidding, but we we we're not sure how to make these 2, like, match. Now this orthogonality, why is it important for safety? So so 1, we we we need to understand that separation because we could have very intelligent beings that are also very nasty. Right? Because the goals are malicious. It's a mistake to think that because you're smart, you're good. What? Because because of this separation. You could have a lot of knowledge, a lot of intelligence to apply that knowledge in any circumstance, which is reasoning and optimization. But what you do with that, like what goal you try to achieve, what values, you know, you you put to decide how to act, can be chosen completely independently. So you can have something very intelligent and with good goals or with bad goals. So in the case of you can think of a tool, for example. Like, think of any tool is generally dual use. Depending on what I choose to do with a tool, I can harm or I can help. Knives, you know, whatever. So that's that's the separation. Now why we could use that to our advantage? Why not build machines that understand the world like a scientist? Not a business person, a scientist, not trying to be a product to cater to our needs, but just be truthful and humble in exactly the right measure. And we could use that without putting the goals part, which is potentially dangerous. We could use that to advance science, to advance medicine, to figure out, you know, cures for diseases, to figure out how to deal with climate change, to figure out how to grow food more efficiently. So this is science. And really science is about understanding, and then using that understanding to answer critical questions that we care about. So we could potentially build machines that are helping us solve the challenges of humanity without taking the

risk of putting this goal seeking machinery into theirs. It doesn't solve all the problems, but at least we know that it's not gonna blow in our face. But a human could still use these things for, you know, designing new weapons, for example. So it doesn't solve the social problem. It doesn't solve the political problem, but at least we don't get this unintentional loss of human control, could spell, you know, catastrophic outcomes.[14]

Tim Scarfe 00:40:39

I watched your your monk debate with Melanie Mitchell, and I think that was the paper clip example. And and she said, why would such a super intelligent machine not know that it's making paperclips and it's doing something really silly? And you're proposing a kind of system where we can stop the AGI from taking control of its goals in a dangerous way. We can have a because[15]

Professor Yoshua Bengio 00:41:04

it doesn't have any goal. It's just trying to be truthful to the data that it's seen and trying to find explanations for the data.

Tim Scarfe 00:41:11

So so a a non agential form of AI that has no goal. That's right. And what would what does that mean in in practical terms? So it doesn't have this feedback loop. That's right. It's like an oracle.

Professor Yoshua Bengio 00:41:21

Yes.

Tim Scarfe 00:41:22

Yes. So A probabilistic oracle because truth,

Professor Yoshua Bengio 00:41:26

you know, is never binary.

Tim Scarfe 00:41:28

Yes. There's uncertainty, and you need to also be accurate about that. Interesting. But we were saying before that the the magic of distributed superintelligence that we're in is this memetic information sharing tool use, culture,

[14] The Superintelligent Will (Orthogonality Thesis) — paper
[15] Munk Debate on AI Existential Risk — tool

Professor Yoshua Bengio 00:41:44

cultural transmission. So would it you know, would we be limiting the intelligence by, you know, using it in this restricted way? Yes. But we might also save ourselves. Indeed. And we could potentially use that, like, non agentic scientist AI to help us answer the most important question, which is how do we build an agentic AI that is safe? Or maybe there is no solution, but at least we would have, like, a super scientist to help us or multiple ones to help us figure out this question. And we would need to figure it out because people are people and they want agents. Yes. But we should just do it carefully. Right now, we are building agents, and we are hoping that these agents will not try to fool us while they help us build the next generation of AI systems. But we're we're we're building on something that may be dangerous. If we construct a ladder of building more and more intelligent systems on top of non agentic series of rungs, At least for that part, we are safe. When we decide to jump the agency

Tim Scarfe 00:43:18

challenge, we might do it in a safe way because we're relying on intelligence, knowledge, understanding that isn't is is truthful, isn't is is trustworthy, is not trying to, you know, do something for itself. It's just trying to be answering the questions. And the questions are things like, you know, would this work or, you know, what what sorts of algorithms would have which properties and so on. How might that change our agency? So, you know, we were saying before that certainly a lot of large distributed systems might take away our agency, but even if we had very sophisticated tools and oracles, in some limiting circumstances they could, you know, really improve someone's agency to do bad things. Of course. The

Professor Yoshua Bengio 00:44:09

non agent, you know, AGI or superintelligent system only solves the problem of loss of human control. And and it doesn't even completely solve it because you could still have a human turn the non agent

International AI Competition and Hardware Governance

Professor Yoshua Bengio 00:44:25

system into an agent. It's easy to turn an oracle into an agent. You just take the current state as input and also add the question, in order to achieve this goal, what should I do? And then you got an agent. And then you take the output as well as what you observe back as a new, you know, additional information for the input. So you can create that loop. When you close the loop, you've got an agent. And, of course, that agent could do could be potentially dangerous. And more importantly, even if it is not dangerous, humans could ask questions that, you know, allow them to gain power and, you know, do bad things and and, you know, take control over other humans or harm people because they have their own, whatever, military goals or political goals or even just the economy goals.[16]

Tim Scarfe 00:45:27

What's your p do?

Professor Yoshua Bengio 00:45:30

I I'm very agnostic about this whole thing. Like, I really don't know. And so I prefer to say that I I have a lot of uncertainty about the different scenarios. What I do know is that the, like, really bad scenarios can have catastrophic consequences, including extinction of humanity, and that there are clear mathematical arguments why some of these scenarios would happen. Now there are so many other things we don't control, like regulation or, you know, advances in technology and so on. It doesn't mean that it's going to happen. Maybe we find fixes. But I think these arguments are sufficiently compelling that they tell me we should take care of that problem, and we have a bit of urgency because we don't know when the current train is going to reach AGI.

Tim Scarfe 00:46:39

Do you have a sense, though? Do have

[16] Can a Bayesian Oracle Prevent Harm? — paper

Professor Yoshua Bengio 00:46:41

a sense of how close we are? Again, I'm very agnostic. Honestly, it could be a few years, like Dario and Sam are saying, or it could be decades. We need to plan for all of these because nobody has a real crystal ball. Maybe the people in the companies have a bit more information. Although different companies contradict each other on this, so I would, you know, take the whole thing with a grain of salt. But from the point of view of policymaking or collective decision about what to do about AI, we need to look at the plausible worst case. Like, if it's very fast, are we ready? Like, do we have the mitigations, the technical mitigations? Do we have even the ways to assess the risks? No. Do we have the social infrastructure, governance, regulation, international treaties to make sure that everywhere we develop AGI, we do it right? No. It's no, no, no. But maybe if it's 20 years, we figure out all these questions, the the political ones and the technical ones. But but right now, we're far from having the answers.

Tim Scarfe 00:48:04

Do you have any ideas? Because you know, like, we're we're in this competitive global landscape, different cultures, different values, and and so on. Yeah. How might we build an effective AI governance system?

Professor Yoshua Bengio 00:48:16

The endgame is 1 where no single person, no single corporation, no single government has too much power. That means it has to be that the governance, like the rules that we decide, you know, how we use AI and so on, have to be, you know, multilateral and involve many countries. And and by the way, it's of course, there's like a couple of or a few countries that are leading and so on. And what would be their interest in sharing that power? Well, because eventually, some other country it's like nuclear proliferation. Like, eventually, some other country will figure it out. And we don't want them to build a monster that kills us. Or we don't want them to build something that allows them to design weapons that kill us. Right? So there are lots of bad scenarios where the only option is somehow we find a way to coordinate internationally. Now on our way there, there are many obstacles. But but if we get to that stage and we have the right technical and governance guardrails, then we we could be in a world where we just reap the benefits and we avoid the catastrophic outcomes. So on our way there, 1 of the obstacle is the competition between The US and China. 1 of the reasons why many people are saying, oh, we can't slow down, you know, to take care of safety is because, you know, that would allow the Chinese to, you know, leap forward. And the Chinese are thinking the same thing, of course. So you can see that's a dangerous game. But it there's also a real danger that other countries

will do something dangerous. And until enough, I mean, of the leading countries understand the existential risks, it's going to be difficult to to to negotiate treaties. And then for those treaties to work, there'll be a need for verification technology. Like, okay. We don't trust each other because the same AGI could be used as a weapon or to build new weapons. Right? So how do I know that behind my back, you're not actually using your AGI for something that would be bad for me? So we need a way or multiple ways to to do these verifications, and and there are researchers working on this. The most promising is what's called hardware enabled governance. The idea is following on existing approaches that already companies are using even in your phones, but in, you know, in in other, like, hardware devices, for example, for privacy reasons and so on. We we already have cryptographic methods to obtain some guarantees about the code that is running in in a chip and so on. And so we could push in that direction and end up with AI chips that can only be used in some ways that have been agreed upon to to to simplify.

Tim Scarfe 00:51:58

Do you remember that piece in time when Elias just hypothetically spoke about bombing data centers? Right. And I mean, of course, that's an extreme example. Right. And maybe we might have like a fire alarm, some way of detecting advanced capabilities being developed. But do you think that we might need to make decisions of that magnitude?[17][18]

Professor Yoshua Bengio 00:52:19

That's a scenario that I can't rule out. Obviously, should try to avoid it, but I can imagine. Actually, 1 version of this is imagine a country that is not leading an AI and has nukes. You can guess which 1. And they don't wanna see, say, us develop weaponry that would be way above what they can defend against. So what's their option?

Tim Scarfe 00:52:57

Yeah. Press the button.

Professor Yoshua Bengio 00:52:58

Destroy destroy our data centers.

[17] Hardware-Enabled AI Governance Memo — paper
[18] Yudkowsky's TIME Article on AI Containment — website

Tim Scarfe 00:53:00

Yes.

Professor Yoshua Bengio 00:53:01

So data centers are going to become a military asset when they can run AGI.[19]

Tim Scarfe 00:53:11

If what you say is true, this is a bit like when we developed nuclear weapons. Yeah. It it it creates this very rapid power imbalance, which has ripple effects. Yeah. That's how you see it.

Professor Yoshua Bengio 00:53:24

But we need to think ahead of these possibilities, even if it's 20 years from now. Know, think how much time it took to sign the nuclear nonproliferation treaties in the sixties, and and the negotiations started right after the end of World War 2. So that's almost 20 years. And that's kind of the timeline where, I would say, there's a high probability that we figure out AGI. Very high probability. Could I play devil's advocate just for a minute, though? There there are people who just think that AI is not really as smart as we think it is, and that these risks are are overblown. What what would you say to those people? I hope they're right. But what what I perceive is that the AIs we're building now have superhuman capabilities on some things, and also subhuman capabilities that, you know, even a child would not, you know, make these mistakes. The other thing I observed is the trend. If I look at the last 10 years and the benchmarks, the old ones and the new ones, that it's very clear. We continue making progress. It's like, you know, there's no stop in sight. Maybe there will be 1. Maybe we hit a wall. I don't know. But if we want to be on the prudent side, we should consider the possibility that we continue that for, you know, a few years, and reach a point where whether you call it AGI or not, reach a point where the capabilities are sufficiently dangerous that in the wrong hands, they could be catastrophic. And eventually, even without having full dominance over all human abilities, if an AI has enough is superhuman in enough areas, it could be dangerous as well. Right? I mean, persuasion is an example. Only need this 1. Right? Persuasion, and you can control people, and then people can do your bidding. So you see, you don't need to be to have an AI that knows everything. It just needs to press our buttons very, very intelligently. I'm just saying that we make a fuss about this concept of AGI, but really from a safety and security point of view, we should be thinking about capabilities, individual capabilities that with the wrong goals, the orthogonal-

[19] Nuclear Non-Proliferation Treaty (1968) — website

ity principle, can become dangerous when against us. Right? So whether it's in the[20]

Tim Scarfe 00:56:07

hands of other humans or an AI that we lost control, we don't want that to happen. Indeed. So you won the the Turing Award with Jeff and and Jan, which is basically the Nobel Prize of computing.

Professor Yoshua Bengio 00:56:20

And you But Jeff got the

Tim Scarfe 00:56:21

real Nobel Prize. Real. I know. I know. I did did think that when I said it. But but you you wrote that you you you feel a sense of loss over the potential negative consequences of your life's work in AI. This is your life's work. I mean, it's it's incredible what you've done. How how do you reconcile that? I'm a human.

Data Centers as Military Assets

Professor Yoshua Bengio 00:56:43

I should have seen it coming. You know, I I I had some students who were worried about this a long time ago, and and they told me about it. And I read a few papers and books and so on, but I I thought, oh, that's great. Some people are worrying about this, and we should have, like, some research to understand those possibilities. I'm glad some people are working on it. Me too. I wasn't taking it seriously for myself until ChatGPT came out. Then I realized that I had a responsibility, and I wouldn't be comfortable with myself if I didn't do everything I can, everything I can, to contribute to reducing the risks. On on the basis that it might be a risk? On the basis that it might be, maybe not. But there's enough indications that it can be a catastrophic risk that I felt I I couldn't do anything else. But pivot, go against my own community. I I I've been among all the other AI people saying, oh, AI is great. It's going to be, you know, bringing so much benefits to society. And I had to change that mental picture to incorporate also the catastrophic risks.

[20] 2018 Turing Award — tool

Tim Scarfe 00:58:14

These ideas do creep up on you. Mean, I've spoken to so many, you know, safety folks, and the ideas just get slowly baked in over time. When I interviewed you last time, I saw that you had a copy of the precipice on your bookshelf. I'm sure that was very influential for you. But how do you think about the zeitgeist of this kind of movement? How how is how is it changing over the last few years? I'm new to this. I'm learning.[21]

Professor Yoshua Bengio 00:58:47

I I I don't think I expected in the last year and a half that I started getting involved with, like, signing those letters and so on and talking to journalists about it. I didn't think that we would have as much impact as we've had. So, you know, the glass is half full. There's much more global awareness of this issue. The half empty issue is that there is the awareness of the risks is extremely superficial, even in the AI community. I I talked to a lot of AI researchers, and then I asked them, so have you been reading or, like, thinking about this discussion and this debate, you know, what really think? And most of the time, I get an answer that tells me they read the headlines. And then maybe they made up their mind, you know, 1 way or the other. But very few people take the time to dig, like read up, think about it, make up their own mind, try to see the logic of different positive or negative scenarios. And that's true for AI scientists. It's also true, of course, in the general population where they don't have the references. Like, they they they think about science fiction templates and politicians. It's it's the same thing.

Tim Scarfe 01:00:28

Is the movement a little bit Western centric? And and if it is, what why is that?

Professor Yoshua Bengio 01:00:35

So I've been talking to people in developing countries. I've been also talking to people in China. And there's it it this it's easy to understand from their point of view that the problem is ours. Like, we're creating the problem. And their problem is being behind. And they're gonna build AI systems that are gonna be weaker than the ones, like, the frontier ones that we build in the West. So their AI systems are not gonna be dangerous. We know that, like, smaller ones, less capable ones are less dangerous. It's all about capability. Like, risk is directly associated with capability. Right? Uh-huh. I mean, risk comes from capability and goals, like intentions. So if you don't have the capability, you can't do a lot of harm. So from their point of view, they wanna reap the benefits, and they

[21] The Precipice: Existential Risk and the Future of Humanity — book

don't wanna be left behind. By the way, that's true of China as well. They feel they are behind. They're just a little bit behind, but if you go to you know, I was in Vietnam recently to get another price. And they're developing quickly. They want to embrace technology, science. But these issues of catastrophic risks, they basically are in the hands of a few Western companies. They they can't do anything about it. They think they can't do anything about it, but they think they can develop their economy by deploying AI, you know, training their workforce to to, like, engineer it in various applications and and building their own even sovereign capabilities. But it's gonna lag for a while.

Tim Scarfe 01:02:36

Yeah. I wonder by how much, because, you know, Alibaba have just released some incredibly strong language models. And Right. I suppose the question is, what what is the moat? Is it technical knowledge, or is it just raw data and compute? All of these things. And capital, which is kind of connected to all 3. Yeah.

Professor Yoshua Bengio 01:02:57

I don't think so if if we were to freeze the scientific and engineering advances in AI, there would be no moat. It would be quickly eaten up. But, of course, that's not the reality. The reality is we're continuing to accelerate towards AGI.

AI Safety Evolution and Regulation ●

Professor Yoshua Bengio 01:03:22

And there's this possibility, at least, of rich get richer. That as we advance the, for example, programming abilities of AI, we can help advance our AI research faster than otherwise. So the companies, for example, that are building the frontier AI, they, you know, they have these these models that they haven't been deployed yet. So they have some number of months that nobody has access to their system except them, and they can use them to design the next generation. Eventually, when we approach AGI, that means we start having AIs that are as good as our best AI researchers. Now there's an interesting thing that happens that's really worth explaining. When you train 1 of these frontier models, let's say it takes a few 100,000 GPUs. I mean, the future ones, more or less, or maybe the current ones that we don't know about yet. But that's the order of magnitude. Once it's trained, you can use the same GPUs to create a few 100,000 copies of the AI all running in parallel. And in fact, more because if you wanna think of them as like people doing a particular task, they can work 24 7. Right? So let's say 1 of these companies is able to build a system that is as good as their 5 best AI researchers,

like the cream of the crop. After this AI has been trained, that's really good at AI research, they go from 5 to 500,000. That's a big jump. In reality, it's gonna be intermediate steps where the AI isn't quite as good as the best ones, but now they're increasing the workforce of, like, you know, different abilities in in the process of creating AI. So it's not gonna be, like, necessarily, like, a sharp turn, but there's a chance that whoever's leading is gonna start leading more because they can use their own AI to to advance. I don't know if that's going to happen, but it's a plausible scenario which has a flavor of winner take all, which the companies are well aware of. Right? And so that's 1 reason why they're racing. If they thought that being second would be good enough, then there wouldn't be as much pressure. But they all think that it's a winner take all game.

Tim Scarfe 01:06:10

Something that Hinton says quite a lot is you could have 1000 AI agents. They could be like Von Neumann, and they're just doing things 1000 times faster. But does it really scale like that? You know, there's this book, The Mythical Man Month, which is that software engineering, it doesn't scale very well. When when you have another person on the team, another person on the team, you get this kind of sharing bottleneck. Do you think a humans. Well, why would it be different?

Professor Yoshua Bengio 01:06:41

Okay. So 1 fundamental difference is the bandwidth. The communication bandwidth between humans is very, very small. A few bits per second. And the communication bandwidth between computers, I don't know the exact numbers, but it's like, you know, 1000000 times or some like many many zeros more. That's a very very good reason why you could paralyze the work a lot more. By the way, that is also the reason why these LLMs know so much more stuff than any of us could. It's it's because you can have 100,000 GPUs each reading a different part of the Internet and then sharing their learning through high bandwidth communication where the weights are shared or the gradients are shared. It's the same process. So the the the kind of collaboration that you're gonna have between computers might be very different from the kind of collaboration we have between humans. It could be much tighter, almost like if it was 1 organism.

Tim Scarfe 01:07:53

Yeah. I I can see the argument. I I have an intuition that the reason why humans struggle to understand each other is we have very situated knowledge and representations. So we understand things very differently. And even with language models, you know, I find that O(one) because it has so many distractors in its context, so it's thinking about this and thinking about this, it gets confused more easily. And in a weird way, even though we've copied the weights of all of these neural networks, because they've taken different trajectories and then they share the information, I'm just speculating here, but it might not be quite as big of an uplift as we think.

Professor Yoshua Bengio 01:08:29

Well, this was an issue 10 years ago. It has been solved to the extent that we can put 100,000 GPUs on a cluster. Yeah. So I'm not saying that the same recipes will work for 1000000 or 10,000,000, but engineers have found ways that you can paralyze very efficiently, at least for training. And, of course, inference is even, like, easier. But in a way, solving a task together is more like training because you need to exchange lots of information to be efficient. So clearly, I don't have the answer to, you know, is that gonna be an obstacle or not? I'm just saying the conditions are quite different. The breakpoint of paralyzation might be very different for that reason. Eventually, yeah, maybe it becomes an obstacle, but but this is so far off our, you know, human experience that it would still be a huge advantage.

Tim Scarfe 01:09:28

How responsible do you think the the hyperscalers are? You know, Dario, for example, he's he's recently become a bit more of an accelerationist. What what what are your kind of perspectives on that?

Professor Yoshua Bengio 01:09:45

I understand the concerns about China. And I but but but I think it's a mistake. I don't think Dario actually makes this mistake. It's a mistake to think that it's either we the West, you know, stays in the lead, and, you know, doesn't deal with safety properly, or we slow down and deal with safety properly, and then maybe China takes over. These are 2 possibilities, but we have enough resources, capital, human, to both do safety right and stay in the lead. The way to to do that is simply to make sure we put enough capital in the safety bucket. Right? Once we understand that humanity's, you know, survival is at stake, it's clearly worth it. Or, you know, once you understand that well, we understand that democracy is at stake, so, you know, you you wanna stay in the lead. You wanna you wanna make sure democracies stay in the lead. And by the way, I think it's also important if you think about democracies being in the lead that it's not just The US being in the lead. Also, for this reason, like, we need to put together all our resources to move towards AGI while doing it safely. So that means we need the capital, not just in The US. We need capital from other democracies. We need the talent from other democracies. We need the energy from other democracies, you know, to run the data centers. We need the electrical grids that maybe, you know, might not be sufficient in The US. So there's a greater chance that we achieve both safety and maintaining sort of democratic advantage if we if we if we take the right decisions and we work with multiple democracies together.

Tim Scarfe 01:12:06

I spClaudeke with Gary Marcus recently, and he was saying that the Silicon Valley companies, it's a little bit like, you know, with cigarettes and social media and so on, and that they're not being sufficiently regulated. I mean, I'll give an example. In the Lex interview of Dario, he was kind of saying that they have these guidelines around reaching certain thresholds of intelligence. And they, of course, make those designations themselves. Now, do loads of loads of great work. On the Claude 1 model, they had Apollo research doing lots of safety engineering and so on. So they do lots of good stuff, but do you think that they should be regulated? Yes.[22]

Professor Yoshua Bengio 01:12:45

It should be obvious. Like, we don't want companies to grade their own homework. We need external neutral evaluations that represent the interest of the public. Now I think the real question is not, you know, should we have regulation or not. It's what regulation? Right? How do we make sure we, you know, we don't stifle these advances? And I think there are answers. So the general prin-

[22] EU AI Act Code of Practice — tool

ciple is don't tell the companies how they should do it, how they should mitigate risks, how they should evaluate risks. Use transparency as the main tool to obtain good behavior. Let me explain why transparency is so powerful. First, the obvious. The companies want to keep a good public image, at least in in democracies. Second, they don't want to be sued. If your, you know, risk assessment becomes a public document, or at least a document that a judge in a court could see, because there there's some national security things, but, you know, some things are gonna be redacted and some things not. But but but presumably a judge could have access to a whole to hold information. So now a judge would have enough information to declare, well, you didn't do as much as you could, given the state of the art in safety, for example. You didn't protect the public, and so now this person or this group of people who, you know, lost billions of dollars are suing you, and they're right. You you you know, you could have done better. So the effect would be obvious, right, that that if if if you know you can be sued because you you act in a dangerous way, then you have to be honest about the risks. Like, as a company, if if you wanna avoid these kind of lawsuits, first, you need to know what risks am I taking, and then how do I control them to balance, you know, the these possibilities. Right? So suddenly, they they have to do the things that we want. I'm not saying this is a perfect process, like but but at least it's an easy 1. So companies should be forced to register. Like, the the the government needs to know what are the big potentially dangerous systems. And then those that register need to tell the government and and the public to the extent, you know, it's it's it's reasonable what their plan is, so called, you know, safety and security frameworks, and what were the evaluations that they did, and then what were the results, and how did they what kind of mitigations do they plan to do, and what kind of mitigations did they actually implement? So if a company says, if we reach that level, we will do x, and then they don't do it. Then then they, you know, they they can be sued if something bad happens. Right? So that's that's how powerful transparency is. And it it doesn't require the state to actually, you know, judge, you know, and tell the companies exactly what to do. Just forces them to disclose all that information, and maybe with independent third parties because the government may not have all the expertise. So we already have companies springing up to do these evaluations so long as they're not paid by the AI company. So we have to be careful. We need to learn the lesson from finance. I think that there's a reasonable path here, which doesn't prevent companies from deciding what is best, both in terms of capabilities and in terms of safety, and stimulates innovation in safety, which is really the thing we need right now.

Tim Scarfe 01:17:22

That sounds quite pragmatic. I mean, was a bit concerned with the FLOPS regulation, and Sarah Hooker did a wonderful paper on that, by the way. But what is the meta game that these companies are playing? Do you remember when Sam Altman went to the Senate, and he was begging them to to regulate him? Should we be cynical about that? Do you think that was just regulatory capture?

Professor Yoshua Bengio 01:17:43

I don't read minds. Worse than that, people can be biased unconsciously because that's what psychologists call motivated cognition. So they might even be sincere. But it's just a story that fits them better, that makes them play a more beautiful role, and we all do that. So by default, I'll assume that, you know, those people are sincere. But because humans can be fooling themselves, they do, we need, like, other eyes on the projects that don't have any financial or personal incentives 1 way or the other, except the, you know, well-being of the public. On the safety front and and maybe this is gonna connect with the questions you wanna ask later. But on the safety front, there are so many open questions. I said we we need to do a lot more research, so that's obvious, but and we need to put the right incentives. But but I want to insist that we we need many different threads of research, many directions. We should welcome all the projects that try to help in evaluations, in mitigation, in even redesigning the way that we build AI. This is so important. In my opinion, it should be like humanity's number 1 project because our future is so much at stake. And so we should put all our minds on figuring out how to to do this safely. And right now, there's a bit of concentration of everybody's doing the same thing or 2 or 3 different things, both on capabilities and on safety. So on capabilities, we see, like, everything doing the same sort of LLMs and RLHF and whatever the recipe is. And and now everybody's gonna be doing, you know, internal deliberation. Right? And on safety, there's also a lack of diversity. We we we really need to invest more broadly. And this is a place where academia can help, because academia is naturally widely exploring. Sometimes academia may not be the right vehicle. If you do a safety project that's also going to have potentially consequences in terms of capability increase, then academia may not be the right bet. Because you you you may not want to publicize advances in capability for reasons similar to why the companies are not publishing their work anymore. I mean, in their case, it's a mix between commercial competition and also being worried about, you know, adversaries using that knowledge against us. So or somebody using that and making a mistake and creating a monster. So there are good reasons why some research needs to be in academia, and some research needs to be in ideally nonprofit organizations. And, of course, the bulk of the research is gonna continue being in industry, but even in safety, like but we need to put the right incentives. Okay. That's what the

preamble. And right now, everyone thinks that in order to build AGI, we need to solve the agency problem. My thesis is, actually, we don't. Like, we can build really useful machines that are not agent, and we can reduce the risks a lot by doing that and still reap a lot of the benefits, and still, you know, not close the door to agency, but doing doing it in a safe way.

Tim Scarfe 01:22:07

Very interesting. And of course, if we want to have academia doing frontier research, they're gonna need billions of dollars.

Professor Yoshua Bengio 01:22:15

Yeah. That's the other problem. And that is also a reason why it'd be good to create an alternative vehicle for AGI research, which is public, nonprofit oriented towards AI that's going to be applied for, you know, dealing with the the biggest challenges of humanity and with safety as sort of the number 1 principle. But that's gonna take multiple governments, billions and billions of dollars. People talk about the of AI. I think that's that's an important part of the picture we should try to paint.

RNNs to Transformers and Back

Tim Scarfe 01:23:00

Yeah. Is there a a clash of incentives as well? Mean, there there are several startups that are really super focused on safety. But just to be profitable, they they have to work on capabilities as well. Is is is that a difficult square to circle? Yes.

Professor Yoshua Bengio 01:23:17

But I think a lot of the safety startups I mean, depends. So some of the safety startups are working on stuff that, like evals, for example, that's not going to increase not going to increase capabilities. Yep. I I imagine, for example, that Ilias Zuskever's startup is is more of the kind you're thinking about.

Tim Scarfe 01:23:36

Yes. Yes, indeed. Well, let's talk about a couple of your technical papers because, mean, to be honest, it's it's overwhelming. You've so many papers just in the last year. But 1 thing that that really jumped out to me is your were RNNs all we needed paper. Could you sketch that out for me?

Professor Yoshua Bengio 01:23:55

When we introduced the attention mechanisms that are currently used in industry and academia, In 2014, actually, it it was using RNNs as the engine. This was pre transformer, which came in 2017. And there's there's an issue with RNNs, way that, like, their normal design is, which is that you can't easily paralyze training over the sequences. So you you have to go once so you got a sequence of, like, words, for example. And so the the the neural net has to process 1 word and the next and the next, and it needs to construct an internal state, that's the recurrent state, from the previous steps in order to feed the next step. The problem if if you just had, like, 1 computer, like, 1 normal computer, the classical CPU, that's fine. But in GPU days where you can paralyze thousand fold, you're like, well, how do I do that? I can't paralyze because I have to do the sequential thing. So we we did a few things at the time, like to parallelize across examples, but but, you know, you lose some some parallelization. And of course, with transformers, you basically use the same architecture except you remove the recurrence, and and now you can do everything in parallel, like for the whole sequence. This is in 1 shot, and you can get the gradients. And so in the last few years, there's been several papers, not just ours, people starting to explore how we could put back a bit of that recurrence into the architecture, because it it has some real advantages. And yeah. So with the right tweaking of the architectures, you can do these things. There are already many possible designs, and we are starting to see these, at least on the small scale, beat the transformers. On the large scale, I I know. I I don't know. But clearly, are some advantages to recurrence.

Tim Scarfe 01:26:00

Yeah. I've got some coming on Friday, by the way.

Professor Yoshua Bengio 01:26:02

Yeah. He's gonna tell you all about it.

Tim Scarfe 01:26:04

Yes. Yes. XLSTM. But in in a way, do do do you think it's a sign that we might have over complicated some of the architectures or gating mechanism, for example? So how how how much was that needed? Oh, I

Professor Yoshua Bengio 01:26:18

don't think so. I think these gates are actually useful. I mean so we did the GRU, which is a simplification of the LSTM a while ago. And so it turns out you can pretty much get rid of, you know, 2 of the gates. But but you still need that non-linearity to get the maximum power. So there's a trade off. There's a trade off. You you lose a bit on the expressive power, but you gain so much in capability because now you can you can train larger models for longer because it's so much faster. Right? So for now, it's beneficial to to do that. That trade off is is is is working.

Tim Scarfe 01:27:02

Very cool. So another paper, a complexity based theory of compositionality. Now, being schooled by Fodor and Paulish and myself, you know, there was there was always this discussion that neural networks, they can't do compositionality. What do you think?

Professor Yoshua Bengio 01:27:21

Well, I think that was a very strong claim that was not supported by anything except intuition. Oh, interesting. Go on. Well, our brain is a neural net. Yeah. What what's the difference? So the difference is that current neural nets, it's not clear how they do symbolic things. And as we said before, right now, the trick is to use the input to output loop to throw in some generation of symbols, like internal deliberation, chain of thought. Right? But it's not completely satisfying, and, clearly, it's not exactly what's going on in the brain. The paper isn't really about architectures, though. Right? It's more about how would we quantify compositionality. I it's it's like a not well defined notion. We we have an intuition. I mean, the experts have an intuition about it. I I think that there are different aspects to it, actually. It's not like a simple thing. And so this paper and other work we're doing is trying to, like, pinpoint with mathematical formula. Like, can we quantify something that would fit our intuitions about compositionality? But in general, a lot of my work in the last few years has been about putting symbolic things in the middle of the computation of neural nets. So these g flow nets and generative flow networks were in general like probabilistic inference machines. So think of neural nets that have stochastic computation, so it's not deterministic. And some of that could be continuous, and some of that could be discrete. So that's where symbols live, in the discrete realm. The problem with these, of course, is that we don't know anymore how to train them. Like with back prop, the usual way doesn't work. And so we've come up with probabilistic inference, amortized inference, you know, chief flow nets, like variational inference, like a bunch of, like, principles and ideas that actually allow to train these kinds of

machineries. And in a way, they are closer to reinforcement learning, where in reinforcement learning, you usually think of a sequence of actions that the agent takes, and they can be discrete. And yet, you are able to, you know, get great gradients. But now think of the same sort of principles or something related where the actions are not in the world, but in your mind. The actions are about computation. What computation should I do next? What deliberation should I do next? In my mind, in order to deliver an answer or prove something or, you know, come up with an explanation. These are the sort of things we'd like to have in neural nets that we don't have to really have the system 2 capabilities.

Tim Scarfe 01:30:47

Yeah. Remember we interviewed that interviewed you about that last time, and my co host, Doctor. Duggar, he likened it to a Galton board. Oh, yeah. You know those things where you put little balls through and you and you kind of like you can tweak. Yeah.

Professor Yoshua Bengio 01:31:00

You can control the probabilities at each step.

Tim Scarfe 01:31:02

Exactly. Yes. Exactly. Very, very good. And it's but that was an alternative to something like Markov chain Monte Carlo, if I remember correctly.

Professor Yoshua Bengio 01:31:12

Yes. Because what these because they're stochastic, really, you can think of them as generative models. They're sampling. But but they're not sampling only, like, at the last step. They're stamp sampling all the way, like, in diffusion

Tim Scarfe 01:31:27

Yes.

Professor Yoshua Bengio 01:31:27

Neural nets. Diffusion neural nets, you've got neural nets that compute something, and then we add noise, and then again and again and again. So it's a stochastic process. Yes. You can also have discrete versions of this, and and GFlow nets are discrete versions of diffusion process. I mean and then you can mix continuous and and discrete. So so they that's actually closer to how the brain works. Like, the brain is stochastic and also has discreteness. So the discreteness is not obvious. The discreteness comes about because the dynamics of the brain when you're becoming conscious of something has contractive properties, just just a mathematical property, which mean that the number of places where you could land this dynamics is now a discrete set. Right? So instead of just a continuous trajectory, you can have arbitrary continuous no. It's not arbitrary. It's

GFlowNets and Compositionality •

Professor Yoshua Bengio 01:32:40

a bunch of trajectories which go to 1 place, a bunch of other trajectories to another place. And so these places, they're, like, symbolic because they, you know, they create a partition of the total set of possible states. And so you're either in this group or in that group or in that group. And the number of these groups is exponentially large, but you get discreetness. So so the brain has both a it's like it has a dual nature, like a from 1 angle, it seems to be just 1 big vector of activations. But from the other, you can read off, like, oh, in which region am I? Oh, that's like this thought, this symbolic compositional object.[23]

Tim Scarfe 01:33:26

Why do we need discreteness?

Professor Yoshua Bengio 01:33:30

That's a good question. Well, clearly, we use it a lot. All of math is basically symbolic. Yes. I mean, even if you manipulate symbols that are about continuous quantities, you get these symbols. So discreteness allows us to construct abstractions. You can think of what we do when we go from a continuous vector space to like a sentence, getting rid of a lot of detail that maybe doesn't matter that much so that we can generalize better. So in particular, you get a lot of this compositionality coming out naturally in discrete spaces, so, know, like in lan-

[23] GFlowNet Foundations — paper

guage. And that is very powerful. That allows us to generalize in ways that may not be as obvious otherwise.

Tim Scarfe 01:34:38

Isn't it fascinating how in the physical world at different levels of scale, you know, in the emergence ladder, you you get this kind of vacillation between discrete and continuous. And perhaps even in the biological world, you see this kind of canalization where at 1 scale things simplify and get compressed, and then they expand again, and then they compress again. Even in neural networks, that's what we do. We expand, we compress.[24]

Professor Yoshua Bengio 01:35:05

You've got lots of discrete phenomena in the real world. You have cell types, for example. You have convergence of behaviors of cells. That's that's 1 that, you know, I looked a little bit at. And in physics, you you've got phase shifts and and phase transitions and things like that. So there's in terms of dynamics, again, when you have contractive dynamics, which means 2 nearby points at the next step get closer, you you get discreteness that, you know, shows up typically. And that happens in many phenomena in nature and in our brain. Before we go, I'm researching an article on creativity, and I'd love

Tim Scarfe 01:35:56

to quote you. What's your definition of creativity? And I know you put a paper out, by the way, you know, which showed that language models can be more creative than than humans. But what what is creativity? That's a good question.

Professor Yoshua Bengio 01:36:11

So I think there are different types of creativities. So to talk about things people see in current AI, you've got the combination of known concepts, creativity, and we're getting pretty good at that with our, you know, state of the art LLMs. Indeed. There's another kind of creativity which is sort of like the new scientific idea. Often, it is a combination of things we know, because we write it, you know, we define it in terms of things we already know. But it's very far out of the things that we've experienced. And I I I suspect that this kind of creativity that's more out of the box is something that requires more of a search type Mhmm. Of computation. So when we do scientific research, there's a kind of a search. We try this, we try that. Of course, our intuition guides us. It's crucial. Right? But it's not like, oh, we have the solution in 1 shot. There's a search. Like in AlphaGo,

[24] Complexity-Based Compositionality Theory — paper

there's a search, and there's intuition. And right now, we haven't reaped the sort of benefits from the search part in our current LLMs and so on.[25]

Tim Scarfe 01:37:50

There's this boundary between combinatorial creativity and and inventive creativity. And I'm I'm not sure whether it's a hard boundary, whether it's a vague soft boundary, but how could we measure this paradigmatically inventive creativity? I don't know.

Professor Yoshua Bengio 01:38:08

I think I think when we see it, we'll recognize it. So if the AI actually makes true discoveries that nobody thought about, I think we'll know we're entering that territory. But that's not that's not like a test you can you can do. But but I do think that at a mathematical level, we can design our methodology so that it will be trying to do that intuition plus search, like system 1, system 2 more. And so I I believe that will deliver, but how do we quantify? So there is a sense in which scientific discoveries are about finding modes, modes meaning like highly probable explanations for the explanations in the space of explanations for how the world works. So there are many possible explanations. Good ones explain the data well. The day when we make a new discovery, we discover a new potential explanation that seems to fit the data well. And we can abstract that into, like, small sized problems as mode discovery in a jargon of, like, probabilistic machinery. Right? So if if an AI is trying to discover all the good things, like all the good explanations, it's gonna be intractable, but it might be more efficient at finding new modes that it didn't know. And some of the tasks that we can design will be focusing on this ability. So I think I think there's a way to answer your question and do it even at a small scale. Like, we don't need to solve AGI for this. We can design algorithms that will be more creative in their little world. Yeah.

Tim Scarfe 01:40:13

I I love this casting of creativity as epistemic foraging because that that gives it an intrinsic value. But there's also this idea potentially that it's a social phenomenon or it's observer relative, you know. So categories, let's say Move 37 was actually something that we came to recognize as a collective as being a thing. And and that's how it works. But I I suppose that there are different ways to think about it.

[25] Discrete Attractor States in Neural Systems — paper

Professor Yoshua Bengio 01:40:38

Yeah. Yeah. I think I think the go moves that we didn't expect were a good a good way to think about that. But I'd like to think also about something a little bit more general and abstract, which is this mode discovery. Epistemic foraging, as you call it. I like that term.

Tim Scarfe 01:40:56

It's it's Carl Friston's term.

Professor Yoshua Bengio 01:40:57

Ah, okay. Well, it's exactly right. It's it's foraging. It's exploration. And you know when you found something good, but you don't know where it is. So how do you guess where good things are in a very high dimensional space? Well, you need to have good intuition. But it needs to be accompanied with a bit of search. By the way, a lot of that search for humans is not happening in individual brains, it's happening at the level of the collective.

Tim Scarfe 01:41:28

Yes, indeed. Professor Bengio, thank you so much for joining us today. It's been an absolute honor. Thank you

Professor Yoshua Bengio 01:41:33

so much. Pleasure. Thanks for having me. Amazing.

References

[1] **AI Risk Statement**
https://www.safe.ai/work/statement-on-ai-risk
paper

[2] **Professor Yoshua Bengio**
https://yoshuabengio.org/
person

[3] **The Bitter Lesson**
http://www.incompleteideas.net/IncIdeas/BitterLesson.html
website

[4] **ARC-AGI Challenge**
https://github.com/fchollet/ARC-AGI
tool

[5] **Transductive Active Learning Framework**
https://arxiv.org/pdf/2402.15898
paper

[6] **Thinking, Fast and Slow (Kahneman)**
https://en.wikipedia.org/wiki/Thinking,_Fast_and_Slow
book

[7] **Designing Agent Incentives to Avoid Reward Tampering**
https://deepmindsafetyresearch.medium.com/designing-agent-incentives-to-avoid-rew
paper

[8] **Agency (Stanford Encyclopedia of Philosophy)**
https://plato.stanford.edu/entries/agency/
website

[9] **Reward Tampering and AI Safety**
https://hdsr.mitpress.mit.edu/pub/w974bwb0
paper

[10] **Sycophancy to Subterfuge: Reward Tampering in LLMs**
https://arxiv.org/pdf/2406.10162
paper

[11] **AI Alignment Failure Modes**
https://arxiv.org/abs/2410.19749
paper

[12] **Deep RL from Human Preferences**
https://arxiv.org/abs/1706.03741

paper

[13] **Instrumental Convergence Thesis**
https://arxiv.org/pdf/2401.15487
paper

[14] **The Superintelligent Will (Orthogonality Thesis)**
https://nickbostrom.com/superintelligentwill.pdf
paper

[15] **Munk Debate on AI Existential Risk**
https://munkdebates.com/debates/artificial-intelligence
tool

[16] **Can a Bayesian Oracle Prevent Harm?**
https://arxiv.org/abs/2408.05284
paper

[17] **Hardware-Enabled AI Governance Memo**
https://yoshuabengio.org/wp-content/uploads/2024/08/FlexHEG-Memo_August-2024.pdf
paper

[18] **Yudkowsky's TIME Article on AI Containment**
https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/
website

[19] **Nuclear Non-Proliferation Treaty (1968)**
https://www.un.org/disarmament/wmd/nuclear/npt/
website

[20] **2018 Turing Award**
https://awards.acm.org/about/2018-turing
tool

[21] **The Precipice: Existential Risk and the Future of Humanity**
https://www.amazon.com/Precipice-Existential-Risk-Future-Humanity/dp/0316484911
book

[22] **EU AI Act Code of Practice**
https://digital-strategy.ec.europa.eu/en/news/meet-chairs-leading-development-fir
tool

[23] **GFlowNet Foundations**
https://arxiv.org/abs/2111.09266
paper

[24] **Complexity-Based Compositionality Theory**
https://arxiv.org/abs/2410.14817

paper

[25] **Discrete Attractor States in Neural Systems**
https://arxiv.org/abs/2302.06403
paper