**Yann LeCun:** 0:00

Even if you train a system to have a world model that can predict what's going to happen next, the world is really complicated and there's probably all kinds of situations that the system hasn't been trained on and need to. You know, fine-tune itself as it goes. The question of how we organize AI research going forward, which is somewhat determined by how afraid people are of the consequences of AI. So if you have a rather positive view of the impact of AI on society and you trust humanity and society and democracies to use it in good ways, then the best way to make progress is to open research.

**Craig Smith:** 0:35

AI might be the most important new computer technology ever. It's storming every industry and literally billions of dollars are being invested, so buckle up. The problem is that AI needs a lot of speed and processing power. So how do you compete without cost spiraling out of control? It's time to upgrade to the next generation of the cloud Oracle Cloud Infrastructure, or OCI. OCI is a single platform for your infrastructure, database, application development and AI needs. OCI has four to eight times the bandwidth of other clouds, offers one consistent price instead of variable regional pricing and, of course, nobody does data better than Oracle. So now you can train your AI models at twice the speed and less than half the cost of other clouds. If you want to do more and spend less, like Uber 8x8, and Databricks Mosaic, take a free test drive of OCI at oracle.com/eyeonai. That's E-Y-E-O-N-A-I all run together oracle.com/eyeonai.

**Craig Smith:**

Hi, I'm Craig Smith. This is Eye on AI. In this episode, I speak again with Yann LeCun, one of the founders of deep learning and someone who followers of AI should need no introduction to. Yan talks about his work on developing world models, on why he does not believe AI research poses a threat to humanity and why he thinks open source AI models are the future. In the course of the conversation we talk about the new model Gaia 1, developed by a company called Wayve.AI. I'll have an episode with Wayve's founder to further explore that world model, which has produced some startling results. I hope you find the conversation with Yan as enlightening as I did.

**Yann LeCun:** 2:42

I mean, first, the notion of a world model is the idea that the system would get some idea of the state of the world and be able to predict the sort of following states of the world resulting from just the natural evolution of the world or resulting from an action that the agent might take. If you have an idea of the state of the world and you imagine an action that you're going to take and you can predict the resulting state of the world, then that means you can predict what's going to happen as a consequence of a sequence of actions. That means you can plan a sequence of actions to arrive at a particular goal. That's really what a world model is. At least that's what the Wayve people have understood the word in other contexts, like in the context of optimal control and robotics and things like that. That's what a world model is. Now there's several levels of complexity of those world models, whether they model yourself, the agent, or whether they model the external world, which is much more complicated. Training a world model basically consists in just observing the world go by and then learning to predict what's going to happen next, or observing the world taking an action and then observing the resulting effect an action that you take as an agent or an action that you see other agents taking. That establishes causality. Essentially, you could think of this as a causal model. Those models don't need to predict all the details about the world, they don't need to be generative, they don't need to predict exactly every pixels in a video, for example, because what you need to be able to predict is enough details, some sort of abstract representation, to allow you to plan. You're assembling something out of wood and you're going to put two planks together and attach them with screws. It doesn't matter the details of which type of screwdriver you're using or the size of the screw within some limits and things like that. There are details that in the end don't matter as to what the end result will be or the precise grain of the wood and things of that type. You need to have some abstract level of representation within which you can make the prediction without having to predict every detail. That's why those JEPA architectures I've been advocating are useful. Models like the Gaia 1 model from Wayve actually make predictions in an abstract representation space. There's been a lot of work in that area for years, also at FAIR, but generally the abstract representations were pre-trained; the encoders that would take images from videos and then encode them into some representation were trained in some other way. The progress we've made over the last six months in self-supervised learning for images and video is that now we can train the entire system to make those predictions simultaneously. We have systems now that can learn good representations of images. The basic idea is very simple. You take an image, you run it through an encoder, then you corrupt that image. You mask parts of it, for example, or you transform it in various ways. You blur it, you change the colors, you change the framing a little bit and you run that corrupted image through the same encoder or something very similar, and then you train the encoder to predict the features of the complete image from the features of the corrupted one. You're not trying to reconstruct the perfect image, you're just trying to predict the representation of it, and this is different. This is not generative in the sense that it does not produce pixels, and that's the secret to getting self-supervision to work in the context of images and video. You don't want to be predicting pixels. It doesn't work. You can't produce pixels as an afterthought, which is what the Gaia system is doing by sticking a decoder on it and with some diffusion models that will produce a nice image. But that's kind of a second step. If you train the system by predicting pixels, you just don't get good representations, you don't get good predictions, you get blurry predictions most of the time. So that's what makes learning from images and video fundamentally different from learning from text, because in text you don't have that problem. It's easy to predict words, even if you cannot make a perfect prediction, because language is discrete. So language is simple compared to the real world.

**Craig Smith:** 7:43

And there's a lot written right now about the energy required in the computational resources GPUs required to train language models. Is it less in training a world model, like using I-JEPA architecture?

**Yann LeCun:** 8:06

Well, it's hard to tell because there is no equivalent training procedure, self-supervised training procedure for video, for example, that does not use JEPA. The ones that are generative don't really work.

**Craig Smith:** 8:21

Yeah, yeah Well, but this architecture could also be applied to language, couldn't it?

**Yann LeCun:** 8:29

Oh yeah, absolutely yeah. So you could very well use a JEPA architecture that makes predictions in representation space and apply it to language.

**Craig Smith:** 8:40

Yeah, definitely, and in that case would it be less computationally intense than training a large language model. It's possible.

**Yann LeCun:** 8:53

It's not entirely clear either. I mean, there is some advantage, regardless of what technique you're using, to making those models really big. They just seem to work better if you make them big. So if you make them bigger, right. So scaling is useful. Contrary to some claims, I do not believe that scaling is sufficient. So, in other words, we're not going to get anywhere close to human level AI, in fact, not even animal level AI, by simply scaling up language models, even multi-model language models that we applied to video. We're going to have to find new concepts, new architectures and I've written a vision paper about this a while back of a different type of architecture that would be necessary for this. So scaling is necessary, but not sufficient, and we're missing some basic ingredients to get to human level AI. We're full by the fact that LLMs are fluent and so we think that they have human level intelligence because they can manipulate language, but that's false and in fact, there's a very good symptom for this, which is that we have systems that can pass the bar exam, but answering questions from text by basically regurgitating what they've learned more or less by road, but we don't have completely autonomous level five start driving cars, or at least no system that can learn to do this in about 20 hours of practice just like any 17-year-old, and we certainly don't have any domestic robot that can clear up the dinner table and fill up the dishwasher attest that any 10-year-old can learn in one shot. So clearly we're missing something big, and that something is an ability to learn how the world works and the world is much more complicated than language and also being able to plan and reason. Basically having a mental world model the world goes on that allows us to plan and predict consequences of actions. That's what we're missing. Let's take a while before we figure this out.

**Craig Smith:** 11:17

You were on another paper that talked about augmented language models and the embodied touring test. Was that the same paper? The embodied touring test? Can you talk about that? First of all, what is the embodied touring test? I didn't.

**Yann LeCun:** 11:40

I didn't quite understand that well, okay, it's, it's a different concept, but it's basically the idea that you it's based on the, on the, the Moravec paradox, right? So Moravec many years ago. Notice that things that appeared Difficult for humans turned out to sometimes be very easy for computers to do, like playing chess, much better than humans, or, I don't know, computing integrals or whatever, certainly doing arithmetic. But then there are things that we take for granted as humans that we don't even consider them intelligent tasks that we are incapable of reproducing with computers. And so that's where the embodied touring test comes in. Like you know, observe what a cat can do, or how fast a cat can learn new, new, new tricks. Or you know how a cat can plan to jump on, you know, a bunch of different furniture to get to the top of Wherever it wants to go. That's an amazing feat that we can't reproduce with robots today. So that's kind of the embodied Turing test, if you want, like you know, can you make a robot that can behave, have behaviors that are indistinguishable from those of animals first of all, and can acquire new ones, the same with the same efficiency as animals. Then the augmented LLM paper is different. It's about how do you sort of minimally change large language models so that they can use tools, so they can To some extent plan actions? Like you know, you need to compute the product of two numbers, right, you just call a calculator and you know you're going to get the product of those two numbers. And LLMs are notoriously bad for arithmetic so they need to do this kind of stuff or do a search, you know, using a search engine or database, look up or something like that. So there's a lot of work on this right now and it's somewhat incremental, like you know. How can you sort of minimally change LLM and take advantage of their current capabilities but still augment them with the ability to use tools?

**Craig Smith:** 15:38

Yeah, and I don't want to get too much into the threat debate, but you know you're on one side. Your colleagues Geoff and Yoshua are on the other. I recently saw a picture of the three of you. I think you put that up on social media Saying how you know you can disagree but still be friends. This idea of augmenting language models with stronger reasoning capabilities and the ability and agency, the ability to use tools, is Precisely what Geoff and Yoshua are worried about. Can you just get it? Why are you not worried about that?

**Yann LeCun:** 16:27

Okay, so first, first of all, this is not necessarily what you're describing, is not necessarily what they are afraid of. They, they, they are alerting People and various governments and others about various dangers that they perceive. Okay, so one danger when a set of dangers are relatively short term. There are things like, you know, bad people will use technology for bad things. What can bad people use powerful AI systems for and when concerned? That you know governments have been worried about and Intelligent agencies encounter intelligence and stuff like that is. You know, could badly intentioned Organizations or countries use LLM to help them? I don't know how to design pathogens or chemical weapons or other things, or set or cyber attacks? You know things like that. Right Now, those partners are not you. Those partners have been with us for a long time and the question is what incremental Help would AI systems Bring to the table? So my opinion is that, as of today, AI systems are not sophisticated enough to provide any significant help for such badly intentioned people, because those systems are trained with public data that is publicly available on the internet and they can't really invent anything. They're going to regurgitate with a little bit of interpolation if you want, but they cannot produce anything that you can't get from a search engine in a few minutes, so that actually that claim is being tested. At the moment, there are people who are actually kind of trying to figure it out, like is it the case that you can Actually do something? You're unable to do something more dangerous with the sort of current AI technology that you can do with a search engine? Results are not out yet, but my hunch is that you know it's not going to enable a lot of people to do significantly bad things. Then there is the issue of things like code generation for cyber, cyber attacks and things like this, and those problems have been with us for years. And the interesting thing that most people should know, like you know also, is disinformation or Attempts to corrupt the electoral process and things like this, and what's very important for everyone to know is that the best countermeasures that we have against all of those attacks currently use AI massively. Okay, so AI is used as a defense mechanism against those attacks. It's not actually used to do the attacks yet, and so now it becomes the question of you know who has the better system? Like other countermeasures, is the AI used by counter, by the countermeasures, but significantly better than the AI is used by the attack system, the AI is used by the attackers, so that you know, the problem is satisfactorily mitigated, and that's what we are now. The good news is that there are many more good guys and bad guys at the minimum. They're usually much more competent, they're usually much more sophisticated, they're usually much better funded and they have a strong incentive to take down the attackers. So it's a game of cat and mouse, just like every yeah, every security that's ever existed. There's nothing new there. Okay, no, nothing quite to do. You knew. Yeah, but okay, but then there is the question of existential risk, right, and this is something that both Geoff and Yoshua have been thinking of fairly recently. So for Geoff, it's only sort of just before last summer that he became he started thinking about this because before he thought he was convinced that the kind of algorithms that we had were Significantly inferior to the kind of learning algorithm that the brain used, and the epiphany he had was that, in fact, no, because looking at the capabilities of large English models that can do pretty amazing things with a relatively small number of neurons and synapses, he said maybe they're more efficient than the brain and maybe the learning algorithm that we use back propagation is actually better than whatever it is that the brain uses. So he started thinking about, like you know what are the consequences, and but that's very recent and in my opinion he hasn't thought about this enough. Yoshua went to a similar epiphany last winter where he started thinking about the long-term consequences and came to the conclusion also that there was a potential danger. They're both convinced that AI has enormous potential benefits. They're just worried about the dangers. And they're both worried about the dangers because they have some doubts about the ability of our institutions to do the best with technology Whether they are political, economic, geopolitical, financial institutions or industrial to do the right thing, to be motivated by the right thing. So if you trust the system, if you trust humanity and democracy, you might be entitled to believe that society is going to make the best use of future technology. If you don't believe in the solidity of those institutions, then you might be scared. I think I'm more confident in humanity and democracy than they are, and whatever current systems that they are. I've been thinking about this problem for much longer, actually Since at least 2014. So when I started fair at Facebook at the time, it became pretty clear pretty early on that deploying AI systems was going to have big consequences on people and society, and we got confronted to this very early, and so I started thinking about those problems very early on. Things like countermeasures against bias in AI systems, systematic bias, countermeasures against attacks, or detection of hate speech in every language these are things that people at the fair worked on and then were eventually deployed To. Just to give you an example, the proportion of hate speech that was taken down automatically by AI systems five years ago, in 2017, was about 20 to 25%. Last year it was 95%, and the difference is entirely due to progress in natural language. Understanding Entirely grew to transformers that are pretrained, self-supervised and can essentially detect hate speech in any language. Not perfectly Nothing is perfect, it's ever perfect but AI is just massively there and that's the solution. So I started thinking about those issues, including existential risk, very early on, in fact, in 2015, early 2016,. Actually, I organized a conference hosted at NYU on the future of AI where a lot of those questions were discussed. I invited people like Nick Bostrom and Eric Schmidt and Mark Schroepfer, who is the CTO of Facebook at the time, Demis Hassabis a lot of people, both from the academic and AI research side and from the industry side, and there were two days a public day and kind of a more private day. What came out of this is the creation of an institution called the Partnership on AI. This is a discussion I had with Demis Hassabis, which was: would it be useful to have a forum where we can discuss, before they happen, sort of bad things that could happen as a consequence of deploying AI? Pretty soon, we brought on board Eric Horvitz and a bunch of other people. We co-founded this thing called the Partnership on AI, which basically has been funding studies about AI ethics and consequences of AI and publishing guidelines about how you do it right to minimize harm. So this is not a new thing for me. I've been thinking about this for 10 years essentially, whereas for Yoshua and Jeff it's much more recent.

**Craig Smith:** 25:04

Yeah, but nonetheless, this augmented AI or augmented language models that have stronger reasoning and agency raises the threat, regardless of whether or not it can be countered to a higher level.

**Yann LeCun:** 25:26

Right, okay. So I guess the question there becomes what is the blueprint of future AI systems that will be capable of reasoning and planning, will understand how the world works, will be able to use tools and have agency and things like that? Right? And I tell you they will not be autoregressive LLMs. So the problems that we see at the moment of autoregressive LLMs are the fact that they hallucinate, they sometimes say really stupid things, they don't really have a good understanding of the world. People claim that they have some simple word model, but it's very implicit and it's really not good at all. For example, you can tell an LLM that A is the same as B and then you ask if B is the same as A and it will say I don't know or no, right? I mean, those things don't really understand logic or anything like that, right? So the type of system that we're talking about that might approach animal level intelligence let alone human level intelligence have not been designed. They don't exist, and so discussing their danger and their potential harm is a bit like discussing the sex of angels at the moment, or, to be a little more accurate, perhaps it would be kind of like discussing how we're going to make transatlantic flight at near the speed of sound safe when we haven't yet invented the turbojet in 1925. We can speculate, but how do we make turbojet safe? It required decades of really careful engineering to make them incredibly reliable and now we can run like halfway around the world with a two-engine turbojet aircraft. I mean, that's an incredible feat. And it's not like people were discussing sort of philosophical questions about how you make turbojet safe. It's just really careful and complicated engineering that no one none of us would understand. So you know, how can we ask the AI community now to explain how AI systems are going to be safe? We haven't invented them yet, yeah, okay. That said, I have some idea about how we can design them so that they have these capabilities and, as a consequence, how they will be safe. I call this objective-driven AI, so what that means is essentially systems that produce their answer by planning their answer so as to satisfy an objective or a set of objectives. So this is very different from current LLMs. Current LLMs just produce one word after the other, or one token which has a board unit. It doesn't matter. They don't really think and plan ahead. As we said before, they just produce one word after the other. That's not controllable. The only thing we can do is see if what they've produced, check if what they've produced satisfies some criterion or set of criteria, and then not produce an answer or produce a non-answer if the answer that was produced isn't appropriate. But we can't really force them to produce an answer that satisfies a set of objectives. So objective-driven AI is the opposite. The only thing that the system can produce are answers that satisfy a certain number of objectives. So what objective would be? Did you answer the question? What other objective could be? Is your answer understandable by a 13-year-old, because you're talking to a 13-year-old? Another would be this, I don't know, terrorist propaganda or something? You can have a number of criteria, like these guardrails that would guarantee that the answer that's produced satisfies certain criteria, whatever they are? Same for a robot, you could guarantee that the sequence of actions that is produced will not hurt anyone. Like you can have very low level guardrails of this type that say okay, you have humans nearby and you're cooking, so you have a big knife in your hand, don't flair your arms, okay, that would be a very simple guardrails to impose, and you can imagine having a whole bunch of guardrails like this that will guarantee that the behavior of those systems would be safe and that their primary goal would be to be basically subservient to us. So I do not believe that we'll have AI systems that can work, that will not be subservient to us, will define their own goals, they will define their own sub-goals, but those sub-goals would be sub-goals or goals that we set them, and we'll not have all kinds of guardrails that will guarantee the safety and we're not going to. It's not like we're going to invent a system and make a gigantic one that we know will have human-level AI and just turn it on and then, from the next minute, is going to take over the world. That's completely preposterous. What we're going to do is try with small ones, maybe as smart as a mouse or something, maybe a dog, maybe a cat, maybe a dog maybe and work our way up and then put some more guardrails, basically like we've engineered more and more powerful and more reliable turbojets. It's an engineering problem.

**Craig Smith:** 31:06

Yeah, yeah, you were also on a paper. Maybe this is the one that talked about the embodied Turing test on neuro-AI. Can you explain what neuro-AI is?

**Yann LeCun:** 31:24

Okay. Well, it's the idea that we should get some inspiration from neuroscience to build AI systems and that there is something to be learned from neuroscience and from cognitive science to drive the design of AI systems. Some inspiration, something to be learned, as well as the other way around. What's interesting right now is that the best models that we have of how, for example, the visual cortex works is convolutional neural networks, which are also the models that we use to recognize images, primarily in artificial systems. There is information being exchanged both ways. One way to make progress in AI is to ignore nature and just try to solve problems in an engineering fashion, if you want. I found interaction with neuroscience always thought-provoking. You don't want to be copying nature too closely, because there are details in nature that are irrelevant and there are principles on which natural intelligence is based that we haven't discovered. But there is some inspiration to have, certainly in your convolutional net, for, inspired by the architecture of the visual cortex. The whole idea of neural net and deep learning came out of the idea that intelligence can emerge from a large collection of simple elements that are connected with each other and change the nature of their interactions. That's the whole idea. Inspiration from neuroscience has been extremely beneficial so far, and the idea of neural AI is that you should keep going. You don't want to go too far. Going too far, for example, is trying to reproduce some aspect of the functioning of neurons with electronics. I'm not sure that's a good idea. I'm skeptical about this.

**Craig Smith:** 33:33

For example, so your research right now, are you, your main focus is on furthering the JEPA architecture into other modalities, or where are you headed?

**Yann LeCun:** 33:50

Yeah, so, I mean, the long term goal is, you know, to get machines to be as intelligent and learn as efficiently as animals and humans. Okay, and the reason for this is that we need this because we need to amplify human intelligence, and so intelligence is the most needed commodity that we want in the world, right? And so we could, you know, possibly bring a new renaissance to humanity if we could amplify human intelligence using machines, which we are already doing with computers, right, I mean, that's pretty much what they've been designed to do. But even more, you know, imagine a future where every one of us has an intelligent assistant with us at all times. They can be smarter than us. We shouldn't feel threatened by that. We should feel like we are, like, you know, a director of a big lab or a CEO of a company that has a staff working for them or people who are smarter than themselves. I mean, we're used to this already. I'm used to this, certainly working with people who are smarter than me. So we shouldn't feel threatened by this, but it's going to empower a lot of us, right, for humanity as a whole. So I think that's a good thing. That's the overall practical goal. If you want right. Then there's a scientific question that's behind this, which is really what is intelligence and how do you build it? And then which is you know, how can a system learn the way animals and humans seem to be learning so efficiently? And the next thing is, how do we learn how the world works? By observation, by watching the world go by, through vision and all the other senses. And animals can do this without language, right? So it has nothing to do with language. It has to do with learning from sensory perceives and learning mostly without acting, because any action you take can kill you. So it's better to be able to learn as much as you can without actually acting at all, just observing, which is what babies do in the first few months of life. They can't hardly do anything, right? So they mostly observe and learn how the world works by observation. So what kind of learning takes place there? So that's obviously kind of self-supervised, right, it's learning by prediction. That's an old idea from cognitive science, and the thing is, you know, we can learn to predict videos. But then we noticed that predicting videos, predicting pixels in video, is so finitely complicated that it doesn't work. And so then came this idea of JEPA right. Learn representations so that you can make predictions in representation space, and that turned out to work really well for learning image features, and now we're working on getting this to work for video and eventually we'll be able to use this to learn world models where you show a piece of video and then you say I'm going to take this action, predict what's going to happen next in the world and you know, which is a bit where the Gaia system from Wayve is doing at a high level. But we need this at various levels of abstraction so that we can build, you know, systems that are more general than autonomous driving.

**Craig Smith:** 37:07

Okay, that's the yeah, and it's my fault so I won't go over the hour, but is it conceivable that someday there will be a model that you may be embodied in a robot that is ingesting video from its environment and learning, as it's? Just continuously learning and getting smarter, and smarter, and smarter?

**Yann LeCun:** 37:45

Yeah, I mean, that's kind of a bit of a necessity, the reason being that you know, even if you train a system to have a world model that can predict what's going to happen next. The world is really complicated and there's probably all kinds of situations that you, you know the system hasn't been trained on and need to. You know, fine tune itself as it goes. So you know, animals and humans do this early in life by plane. So play is a way of learning your world model in situations that basically you won't hurt you, and but then during life, of course, you know, when we learn to drive, there's all kinds of these mistakes that we do initially, that we don't do after having some experience, and that's because we're fine tuning our world model to some extent. Yeah, learning a new task, we're basically just learning a new version of our world model, right? So, yeah, I mean, this type of continuous, continual learning is going to have to be present, but the overall power and intelligence of the system will be limited by, you know, how much a co-be governor on the internet is using and various other constraints. You know, computational constraints basically.

**Craig Smith:** 39:05

You know you're still young and this. I'm not sure about that. Well, you're younger than Jeff. Let me put it that way.

**Yann LeCun:** 39:14

I'm younger than Geoff, I'm older than Yoshua, yeah.

**Craig Smith:** 39:19

But this, the progress you've made on world models is fairly rapid from my point of view watching it. Are you hopeful that within your career you'll have embodied robots that are building world models through their interaction in reality and then being able to? Well, I guess the other question on world models: do you then combine it with a language model to do reasoning, or is the world model able to do reasoning on its own? But are you hopeful that in your career you'll get to the point where you'll have this continuous learning in a world model?

**Yann LeCun:** 40:07

Yeah, I sure hope so. I might have another, you know, 10 useful years or something like this in research before my brain, you know, turns into bechamel sauce, but, or something like that. You know 15 years, if I'm lucky, so, or perhaps less, but yeah, I hope that there's going to be breakthroughs in that direction during that time. Now, whether that will result in the kind of artifact that you're describing you know robots that can Like you know domestic robots, for example, or sort of in cars that can run fairly quickly by themselves, I don't know, because there might be all kinds of obstacles that we have not envisaged that may appear on the way. No, it's a constant in the history of AI that you have some new idea and a breakthrough and you think that's going to solve all the world's problems, and then you're going to hit a limitation and you have to go beyond that limitation. So it's like you know you're climbing a mountain. You find a way to climb the mountain that you're seeing and you know that once you get to the top you will have the problem solved, because now it's, you know, a gentle slope down and once you get to the top, you realize that there is another mountain behind it that you hadn't seen. So that's been the history of AI right, where people have come up with sort of new concepts, new ideas, new ways to approach AI, reasoning, whatever perception, and then realize that their idea basically was very limited. And so you know this. Inevitably we're trying to figure out what's the next revolution in AI. That's what I'm trying to figure out, and so you know, learning how the world works from video, having systems that have world models, allow systems to reason and plan. And there's something I want to be very clear about, which is an answer to your question, which is that you can have systems that reason and plan without manipulating language. Animals are capable of amazing feats of planning and also, to some extent, reasoning. They don't have language, at least most of them don't and so many of them don't have culture because they are mostly solitary animals. So you know, it's only the animals that have some level of culture. So the idea that a system can plan and reason is not connected with the idea that you can manipulate language. Those are two different things. It needs to be able to manipulate abstract notions, but those notions do not necessarily correspond to linguistic entities like words or things like that. We can have mental images. If you want to think like you do, ask a physicist or mathematician you know how their reason is very much in terms of sort of mental models, nothing to do with language Then you can turn things into, into language. But that's a different story. That's the second, second step. So you know, we're going to have to figure out how to do this. Reasoning, hierarchical planning in machines reproduce this first and then, of course, you know, sticking language on top of it will help, like it will make those systems smarter and be able you know, it will allow us to communicate with them and teach them things and they're going to be able to teach us things and stuff like that. But this is a different question really, the question of how we organize AI research going forward, which is somewhat determined by how afraid people are of the consequences of AI. So if you have a rather positive view of the impact of AI on society and you trust humanity and society and democracy is to use it in good ways, then the best way to make progress is to open research and for the people who are afraid of the consequences, whether they are societal or geopolitical, they're putting pressure on governments around the world to regulate AI in ways that basically limit access, particularly of open source code and things like that, and it's a big debate at the moment. I'm very much on the side, so he's met up very much on the side of open research.

**Craig Smith:** 44:47

Yeah, actually that was something I was going to ask you and now that you brought it up, because I've been talking to people about this and there is a view that, aside from the risks of open source again Geoff Hinton saying, would you open source thermonuclear weapons? Aside from that, the question is whether open source can marshal the resources to compete with proprietary models and because of the tremendous resources required for when you're scaling these models. And there's a question as to whether or not Meta will continue to open source future versions of Llama or not continue to open source, but whether it'll continue to invest the resources needed to push the open source models. So what do you think about that?

**Yann LeCun:** 45:55

Okay, there's a lot to say about this, Okay. So first thing is there's no question that Meta will continue to invest the resources to build better and better AI systems because it needs it for its own products. So the resources will be invested. Now the next question is will we continue to open source the base models? And the answer is probably yes, because that creates an ecosystem on top of which an entire industry can be built, and there is no point having 50 different companies building proprietary, closed systems when you can have one good open source base model that everybody can use. It's wasteful and it's not a good idea. And another reason for having open source models is that nobody has, no entity as powerful as it, thinks it is, has a monopoly on good ideas. And so if you want people we can have good, new, innovative ideas to contribute, you need an open source platform. If you want the academic world to contribute, you need open source platforms. If you want the startup world to be able to build customized products, you need open source base models, because they don't have the resources to build, to train large models. And then there is the history that shows that for foundational technology, for infrastructure type technology, open source always wins. It's true of the software infrastructure of the internet. In the early 90s and mid 90s there was a big battle between Sun Microsystems and Microsoft to deliver the software infrastructure of the internet operating systems, web servers, web browsers and various server side and client side frameworks. They both lost. Nobody is talking about them anymore. The entire world of the web is using Linux and Apache and MySQL and JavaScript, and even the basic core code for web browsers is open source. So open source won by a huge margin. Why? Because it's safer, it gathers more people to contribute. All the features are unnecessary, it's more reliable, Vulnerabilities are fixed faster and it's customizable. So anybody can customize Linux to run on whatever hardware they want. So open source wins.

**Craig Smith:** 48:42

But it's the same thing.

**Yann LeCun:** 48:44

It's going to be the same thing. It's inevitable. The people now who are climbing up like open AI. Their system is based on publications from all of us and from open platforms like PyTorch. ChatGPT is built using PyTorch. PyTorch was produced originally by Meta. Now it's owned by the Linux Foundation. It's open source. They've contributed to it. By the way, their LLM is based on transformer architectures invented at Google. All the tricks to train, all those things came out of various papers from all kinds of different institutions, including academia. All the fine-tuning techniques are the same. So nobody works in a vacuum. The thing is, nobody can keep their advance and their advantage for very long if they are secretive.

**Craig Smith:** 49:41

Yeah, except that with these models, because they're so computer intensive and they cost so much money to train, you need somebody like Meta who's going to be willing to build them and open source them. That's why, when I was asking whether they'll continue, obviously Meta will continue building resource intensive models, but the question is whether they'll continue to open source them.

**Yann LeCun:** 50:14

I'm telling you the only reason why Meta could stop open sourcing models is legal. So if there is a law that adds laws to open source AI systems above a certain level of sophistication, then of course we can do it. If there are laws that, in the US or across the world, make it illegal to use public content to train AI systems, then it's the end of AI for everybody, not just for open source, or at least the end of the type of AI that we are talking about today. We might have new AI in the future, but that doesn't require as much data. And then there is liability. If you believe that someone is doing something bad with an AI system that was open sourced by Meta, then Meta is liable. Then Meta will have a bigger sense of not releasing it, obviously. So the entire question about this is around legal reasons and political decisions.

**Craig Smith:** 51:26

But on the idea of open source winning, don't you need more people or more companies like Metal building the foundation models and open sourcing them? Or could it be an open source ecosystem win based on a single company building the models?

**Yann LeCun:** 51:44

No, I mean you need two or three, and there are two or three, right. I mean, there is this Hugging Face. There is Mistral in France, who is also embracing open source LLM they're very good LLM. It's a small one, but it's very good. There are academic efforts like LAION. They don't have all the resources they need, but they collect the data that is used by everyone, so everybody can contribute. One thing that I think is really important to understand also is that there is a future in which I described earlier, in which every one of us, every one of our interactions with the digital world, would be mediated by an AI assistant, and this is going to be true for everyone around the world, right? Everyone who has any kind of smart device. Eventually, it's going to be in our augmented reality glasses, but for the time being, in our smartphones. And so imagine that future where you are, I don't know, from Indonesia or Senegal or France and your entire digital diet is done through the mediation of an AI system. Your government is not going to be happy about it. Your government is going to want the local culture to be present in that system. It doesn't want that system to be closed sourced and controlled by a company on the west coast of the US. So just for reasons of preserving the diversity of culture across the world and not having our entire information diet being biased by whatever it is that some company on the west coast of the US thinks there's going to need to be open source platforms and they're going to be predominant in at least outside the US for that reason, including China. There is all those talks about oh what if China puts their hand on our open source code? I mean, China wants control over its own LLM because they're a citizen to have access to certain types of information. So they're not going to use our LLMs, they're going to train theirs. That they already have. And nobody is particularly ahead of anybody else by more than about a year.

**Craig Smith:** 54:13

And China is pushing open source. I mean, they're very pro open source within their ecosystems.

**Yann LeCun:** 54:20

Some of them. There is no unified opinion there, but I mean it's the same. In the West, right, there are some governments that are too afraid of the risks and then, or are thinking about it, and some others that are all for open source because they see this as the only way for them to have any influence on the information, the type of information and culture that would be mediated by those systems. So it's going to have to be Wikipedia, right? Wikipedia is built by millions of people who contribute to or from all around the world, in all kinds of languages, and it has a system for vetting the information. The way AI systems of the future will be taught and will be fine tuned will have to be the same way. It will have to be quite sourced, because something that matters to a farmer in Southern India is probably not going to be taken into account by the fine tuning done by some company on the West Coast of the US.

**Craig Smith:** 55:30

AI might be the most important new computer technology ever. It's storming every industry and literally billions of dollars are being invested, so buckle up. The problem is that AI needs a lot of speed and processing power. So how do you compete without cost spiraling out of control? It's time to upgrade to the next generation of the cloud Oracle Cloud Infrastructure, or OCI. OCI is a single platform for your infrastructure, database, application development and AI needs. OCI has four to eight times the bandwidth of other clouds, offers one consistent price instead of variable regional pricing. And, of course, nobody does data better than Oracle. So now you can train your AI models at twice the speed and less than half the cost of other clouds. If you want to do more and spend less, like Uber, 8x8 and Databricks Mosaic, take a free test drive of OCI at oracle.com/eyeonai. That's E-Y-E-O-N-A-I all run together oracle.com/eyeonai.

That's it for this episode. I want to thank You for his time. If you want to read a transcript of this conversation, you can find one on our website ion AI that's eye-on.ai. And remember the singularity may not be near, but AI is changing your world, so best pay attention.