Do LLMs Understand? AI Pioneer Yann LeCun Spars with DeepMind's Adam Brown.

75 MIN · YOUTUBE · HTTPS://YOUTU.BE/YKFQD1_WPBQ
https://youtu.be/ykfQD1_WPBQ

SUMMARY

Adam and Jan discuss the current state and future potential of AI, particularly focusing on large language models (LLMs) and their limitations compared to human intelligence. They explore the technical underpinnings of neural networks, the emergence of deep learning, and the challenges in achieving true understanding and consciousness in AI systems.

- *Jan LeCun emphasizes that neural networks are inspired by human brains but are much simpler and lack true understanding.*
- *The conversation highlights the distinction between LLMs' ability to process language and their lack of genuine comprehension or consciousness.*
- *Adam notes that while LLMs excel in specific tasks, they require vast amounts of data and are less sample efficient than humans.*
- *Both agree that current AI systems are not autonomous and lack the ability to understand the real world as humans do.*
- *Jan expresses skepticism about LLMs achieving human-level intelligence, suggesting that new architectures are needed for true understanding.*
- *Adam remains optimistic about the potential of LLMs, citing their rapid advancements and capabilities in various domains.*
- *The discussion touches on the philosophical implications of AI consciousness and the ethical considerations surrounding AI development.*
- *Jan warns against the over-reliance on LLMs, advocating for diverse AI systems to prevent monopolization of information and ensure safety.*

[music] [music] [music] >> It's a pleasure to have Adam, my colleague and friend, and Jan, who's been with us before. Jan, you really are all over the news right now. Um I've gotten so many people forwarding articles about you this week. It all kicked off on Wednesday. Do you want to discuss the I can just say the headline.

The headline was the equivalent of Jan LeCun, chief scientist, leaves Meta. Um do you care to comment? I can neither confirm nor deny. Okay. [laughter] So all of the press uh the the core that's here to get the scoop cannot get the scoop tonight. >> [laughter] >> All right. Well, um you can come afterwards and buy Jan a drink and see how far you get with um I already had one, but that was a >> [laughter] >> The Frenchman had some wine upstairs.

So we have this era where every time any of us turn on the news, look at the computer, read the paper, we're

confronted with conversations about the societal implications of AI and whether it's about economic upheaval or um the potential for political manipulation or AI psychosis. There's a lot of pundits out there are discussing this and I and and I it is a very important issue. I kind of want to push that towards the end of our conversation because what a lot of people who are discussing this don't have is the technical expertise that's on this stage. And so I really want to begin by grounding this in that technical scientific conversation. And so I want to begin with you, Jan, about neural nets. Here's this instance of kind of biomimicry where you have these computational neural networks that are emulating human networks. Can you describe to us what that means that a machine is emulating human neural networks? Well, it's not really mimicry.

It's more inspiration. The same way, I don't know, airplanes are inspired by by birds, right? The underlying >> work, [clears throat] I thought. Say again? >> But I thought that didn't work. Copying birds with airplanes. Well, in the sense that, you know, airplanes have wings like birds and they generate lift by propelling themselves through the air, but then the analogy stops stops there. And the wing of an airplane is much simpler than the wing of a bird, but yet the underlying principle is the same. So neural networks are a bit like like that, like are like, you know, our two real brains as airplanes are to birds.

They're much simplified in many ways. Um but perhaps some of the underlying principles are the same. We don't actually know because we don't really know the sort of underlying algorithm of the cortex, if you want, or the the the method by which the brain organizes itself and and learns. So we invented substitutes. Um sort of like, you know, birds flap their wings and not airplanes, right? They have propellers, so or or turbojets.

You know, in in neural nets we have uh learning algorithms and they they allow artificial neural nets to learn in a way that we think is similar to how the brains learn. So the brain is a network of neurons. The neurons are interconnected with each other and the way the brain learns is by modifying the efficacy of the connections between the neurons. And the way a neural net is trained is by modifying the efficacy of the connections between those simulated neurons.

Um each of those is like a we call it a parameter. You you you see this in the press, the number of parameters of a neural net, right? So the the biggest neural net at the moment have you know, hundreds of billions of parameters, uh if not more, and um those are the individual coefficients that are modified by by by training. So And how is deep learning uh emerge in this discussion? Cuz deep learning came along the path after thinking about neural nets. And this has been since the '80s or earlier even. Um yeah, no, '80s roughly. Um So early neural nets uh the the first ones that were capable of of learning or learning something useful at least, you know, in the '50s uh were shallow. You could you could basically train a a single layer of neurons, right? So you would feed the input and train the system to produce a particular output and and you could use those things to kind of recognize or or classify relatively simple patterns, uh but not really sort of complex things. And people at the time, even in the '60s, realized that the way to make progress was going to be able to train neural nets with multiple layers. They built neural nets with multiple layers, but they couldn't train all the layers. It would only train the last layer, for example.

Um and they didn't really find uh until the 1980s, nobody found really a good way to train those those

multi-layer systems uh mostly because the neurons that that they had at the time were the wrong type. Um they had neurons that were binary. So neurons in the brain are binary. They they either fire or they don't fire. Um and people wanted to reproduce that. So they they built simulated neurons that would either be active or inactive.

And it turns out for the modern learning algorithms to work, we call them back we call it back propagation, you need to have neurons that have sort of graded responses. Um and uh that only became practical, possible, or people realized it could work in the 1980s. People had the idea before, but they never could really make it work. And so that caused um a renewal of interest in neural nets in the 1980s. They had been largely abandoned in the late '60s and then they came to the fore again in the mid to late '80s. That's when I started kind of my graduate school basically in 1983 and uh there was a wave of interest that lasted about 10 years and then interest went again uh in the mid-'90s until the late 2000 when we rebranded it into deep learning.

Neural net had kind of a bad rep. Um people in computer science and engineering thought neural nets were kind of a bad thing. They had a bad reputation. And so we we branded it into deep learning and sort of brought it back to the fore and then the results were were there in computer vision, in natural language understanding, speech recognition to really convince people that this was uh a good thing. Now, Adam, you at a very young age were interested in theoretical physics, not specifically computer science, and you're watching some of this unfold in some sense from afar. What's the catalyst that sweeps up so many people decades later? There's there's this time where it's of great interest, there's great success in handwriting recognition or in uh uh visual recognition and these things, but it's not sweeping up the world. What what happens that brings us to this point where we're all now talking about large language models?

So many physicists in the last years have pivoted, should we say, from working on physics to working on AI. And it really traces back to some of the work that Jan and others did to prove that it works. Like when it wasn't working, it was just this this thing that's over there in computer science and like of of many things in the world that are not particularly uh that maybe interesting, but not many physicists are paying attention to it. But then after, you know, Jan and some of the other pioneers of this field proved that it would work, it became a totally fascinating subject for physics. That you link up these neurons together in a certain way and suddenly you get emergent behavior that didn't exist at the individual neuron level. That seems like a a subject that physicists who spend their life imagining how the sort of rich pageantry of the world could emerge from simple laws that immediately attracted the attention of many physicists. And nowadays it's a a a very common career path to do a PhD in physics and then apply it to a emergent system. But the emergent system is an emergent network of neurons that collectively give rise to intelligence. Mhm. Now, let's [clears throat] do a lightning round because you raised the dreaded word intelligence.

>> [laughter] >> Um everybody in this room very likely has interacted with something that we're now calling an AI. These are all large language models. And before I ask you to define those for us, I just want to kind of do a lightning round of um of what's your yes or no response to certain things. So um Adam, >> [laughter] >> yes or no, are these AIs, these large language models, uh understanding the meaning of the conversations they're having with us?

Yes or no? Yes. Jan? Sort of. >> [laughter] >> Um perfect. Jan's neurons are not stuck with binary values, is it? >> [laughter] >> Right, exactly. Um it was my fault for giving you a binary choice. Okay, so that allows me to ask the next question cuz it's not a foregone conclusion. If you don't say yes to that, it's going to be interesting what you say to this. Are these AIs conscious? Absolutely not.

Adam? Probably not. Okay. [laughter] Um will they soon be? I think they'll one day be conscious if if progress continues in the way that we're we're continuing. When is hard to say, but One day. Mhm. Yes, for appropriate definitions of consciousness. >> Yes, okay. Well, we do have some philosophers in the house and um we're we're not going to indulge in philosophical definitions of consciousness, or there our hour would go >> [laughter] >> and we'd still be here.

Oh, I just heard that groan, I think from our friends up in the balcony. >> [laughter] >> Um but I have one other question. Okay, no, I have two I have two in the lightning round. Uh are we on the precipice of doomsday or a renaissance in human creativity, Jan? Renaissance. Adam? Most likely renaissance. >> [laughter] >> Um I have to throw this out the same question to the audience, but I'm going to phrase it more colorfully, which I think they'll relate to. Will the robot overlords rise up against humanity? Yes, hands up.

Oh, interesting. Okay, no hands up. Okay, how many robots in the audience? Hands up. Okay. So, okay. So, that's interesting. See, that's cool. It was a little more nose, maybe. Although, the light is blinding. All right, we're going to come back and ask that again at the end. Um so, here we are. These neural nets have been taught to uh execute a process we now call deep learning and there's other kinds of learning that take off. And what are the large language models specifically, which is really what has swept up the news and people's personal experience?

We're we're mostly relating to large language models. And and what are the large language models, Adam? Maybe you could take that. Yeah, so large language model is uh you've probably played with some of them, chat GPT, Gemini made by my company, uh various others um made by other companies. It is a special kind of neural network that's trained on particular inputs, on particular outputs, and trained in a particular way. So, it is at At heart, it is mainly the kind of deep neural network that was pioneered by by Jan and by others, but uh with a particular architecture designed for the following task. Uh it takes text in, so it'll it'll read some uh the first few words of some sentence or the first few paragraphs of some book, and it will try and predict what the next word is going to be.

And so, you take a deep neural network with a particular architecture, and you have it read, basically to first approximation, the entire internet. And for every word that comes along on the entire internet, all of the text data and all other kind of data you can find, uh you then ask it, "What do you think the next word's going to be? What do you think the next word's going to be?" And to the extent that it it gets it right, you give it a little bit of uh reward and strengthen those neural pathways. To the extent that it gets it wrong, you you diminish those neural pathways. And if you do that, uh it'll just start off spewing just completely random words for its prediction, but uh if you train it on a million words, it'll still be spewing random words. If you train it on a a billion words, it'll maybe have just started to learn subject-verb-object and various bits of sentence structure. Uh and if you train it, as we do today, on on a trillion words or more, tens of trillions of words, uh then it'll start

become the conversation partner that you you probably, I hope, uh played around with today.

No, um it it it strikes me as intriguing, like it's it's it amuses me sometimes people get really outraged at their chatbot that they're engaged with when it leads them astray or lies to them. And sometimes I've said, "Well, it's it's doesn't need to be words. It it might as well be colors or symbols. It's just playing a mathematical game, and therefore doesn't have a sense of meaning." Now, I know Adam sort of objected to my summary of that.

Do you think that they are extracting meaning um in the same sense that we do when we are engaging in composing sentences? Well, they're certainly extracting some meaning, um but it's it's a lot more superficial than what most humans uh would extract from from text. Most humans uh intelligence is linked to is is grounded into underlying reality, right? And language is a way to express phenomena or things in that or concepts grounded in that reality. Um and LLMs don't have any notion of the underlying reality. And so, their understanding is is relatively superficial.

Um they don't really have common sense in the in the way that we understand it. Um but if you train them long enough, they they will answer correctly most questions that people will think about asking. That's the way they're trained. You you you collect all the questions that everybody has ever asked them, and then you train them to produce the correct answer for this. Now, there's always going to be new questions or new prompts, new sequences of words for which the system has not really been trained, and for which it might produce complete nonsense. Okay, so in that sense, they don't have the real understanding of the underlying reality, or they do have an understanding, but it's it's superficial.

Um and so, you know, the next question is, how do we fix that? So, I I could play devil's advocate and say, "Well, how do I know that what a human being doing is doing is that much different?" Right? We're trained on lots of language. We get some dopamine hit or some reward system for having said the right word at the right time and the right grammatical structure for the language that we're immersed in. And um and we back propagate.

>> [laughter] >> We try to do a better job the next time. In some sense, how how is that different uh than what a human being is doing? And you're you're saying maybe it's the sensory experience of being immersed in the world? Okay. Um a typical LLM, as Adam mentioned, is trained on tens of trillions of of words, typically It's only a few hundred thousand words, I think. You're just saying sentences. It's combinations. >> 30 trillion 30 trillion words is is a typical size for the training set pre-training of of an LLM.

Uh a A word is represented actually as sequences of tokens, doesn't really matter. Uh and a token is about 3 bytes, so the total is about 10 to the 14 bytes, right? One with 14 zeros um of training data to train those LLMs. And that corresponds to basically all the text that is uh publicly available on the internet plus some other stuff. And it would take any of us something like half a million half a million years for any of us to read through that material, right? So, it's an enormous amount of textual data.

Now, compare this with what a child uh perceives uh during the first few years of life. Um psychologists tell us that a 4-year-old has been awake a total of 16,000 hours. Um and there's about 1 byte per second going through

the optic nerve, every single fiber of the optic nerve, and we have 2 millions of them. So, it's about 2 megabytes per second getting to the visual cortex. Um During 16,000 hours, do the arithmetics, and it's about 10 to the 14 bytes. A 4-year-old has seen as much visual data as the biggest LLM trained on the entire text ever produced.

And so, what that tells you is that there is way more um information in the real world, but it's also much more complicated. It's noisy, it's high-dimensional, it's continuous, and basically the methods that are employed to train LLMs do not work in the real world. That explains why we have LLMs that can pass the bar exam or solve equations or compute integrals like college students and solve math problems, but we still don't have a domestic robot that can, you know, do the chores in the house. We don't we don't even have level five self-driving cars. I mean, we have them, but we cheat.

So, um I mean, we certainly don't have self-driving cars that can learn to drive in 20 hours of practice like any teenager, right? So, obviously, we're missing something very big to get machines to the level of human or even animal intelligence. Well, let's not talk about language. Let's talk about how a cat is intelligent or dog. Um we're not even at that level with AI systems. Adam, you you you impart more comprehension on uh the part of the LLMs at this point already. Uh I think that's right. So, I mean, Jan is making sort of excellent points that the LLMs are much less, for example, sample efficient than humans. Humans or indeed your cat or just a a cat, I don't know if it was your cat or any cat in your >> cat.

>> in your example, um is able to learn from many fewer examples than a large language model, for example, can learn from. It takes way more data to teach it to the same level of proficiency. Um And and that's true, and that that is a thing that is better about uh the, you know, architecture of animal minds compared to these artificial minds that we're building. Um on the other hand, sample efficiency isn't everything. Um we see this frequently, in fact, when we try and, you know, before large language models, when we and put uh machines on you make artificial minds to do other tasks, even the famous chess bots that we built on built on top of large language models, the way they were trained sort of Alpha Zero and various other ones, they would play each other they would play itself at chess a huge number of times and to begin with it would just be making random moves and then every time it won or lost the game when it was playing itself it was sort of uh reward that neural pathway or punish that neural pathway and they play each other at chess again and again and when they played as many games as a human grandmaster has played, they were still making essentially random moves.

But they didn't just we're not confined to making the same number of playing the same number of games that a human grandmaster could play because silicon chips are so fast, because we can build them with such parallel processing, they were able to play many more human many more games than any human could ever play in their lifetime. And what we found is that when they did that, they reached and then far surpassed the level of human chess players. They're less sample efficient but that doesn't mean they're worse at chess. It is clear that they're much better at chess. So too with understanding.

When we it is it is true that we can you know it is harder with these things to you need more samples to get them up to the same level of proficiency but the question is once they've reached that, can we use the fact that they are so much more general and so much more so much faster and more inherent to push beyond that. I mean

another example perhaps with the cat is a cat is in fact even more sample efficient than a human.

A human takes a a year to learn to to walk. A cat learns to walk in a in a week or so. You know, it's much faster. That does not mean that a cat is smarter than a human. It does not mean that a cat is smarter than a large language model. The final question at the end should be what is the capabilities of these things. How far can we push the capabilities and on almost every except for the somewhat impoverished metric of sample efficiency, on every metric that counts, we've pushed these large language models far beyond the frontier of cat intelligence. So >> [laughter] [gasps] >> Yes.

I don't understand why we're not making cats but >> [laughter] >> Sorry, what was I on? I mean certainly the LLMs in question have much more accumulated knowledge than cats or even humans for that matter. And we do have many examples of computers being far superior to humans in a number of you know different tasks like playing chess for example. That's something I mean it just means that humans just suck at chess. That's all it means.

No, we really suck at chess and go by the way even even more. And and many other tasks that computers are much better than than us um at at at solving. Um so certainly LLMs can accumulate huge amount of of of of knowledge and some form of them can be trained to translate languages, understand spoken language and and translate it into another one from you know a dozen languages to another dozen languages in any direction. No human can do this.

Um so they they do have superhuman capabilities but the ability to learn quickly, efficiently, to apprehend a new problem that we've never been trained to solve and be able to come up with a solution um and to really you know understand a lot about how the how the world behaves that is still out of reach of AI systems at the moment. Mhm. I I would I mean we've had recent successes with this where it is not the case that they're just taking problems that they've seen before letter for letter and looking up the answer in a in a lookup table or even that they're uh they are they are in some sense doing pattern matching but they're doing pattern matching at a sufficiently elevated level of abstraction that they're able to do things that if they've never seen before and no no human can do. So there's a there's a competition uh each year called the International Math Olympiad.

Um it is the very smartest uh finishing high school maths uh teenagers in the entire world. They're all given six problems uh each year. The pinnacle of human intelligence. I have some mathematical ability. I look at these problems. I don't even know where to start. Um you know, this this year we fed them into our machine uh as as did a number of other LLM companies and they took these problems they'd never seen before. They were completely fresh, didn't appear anywhere in the training data, completely made up, took a whole bunch of different ideas, combined the different ideas and got a score on these tests that was better than all except the first dozen the top dozen humans on the planet. I think that's that's pretty impressive intelligence. Mhm.

I I guess the question is um back to this idea do they understand? Do you we can look at the mathematics of the model. There's some input data. We understand what it's doing. It is a black box which is kind of fascinating. It's just so complex that it's not as though we can't do that with the human mind either. It's not as though you can

look at the inner workings and and see exactly what they're doing. To some extent it is a black box but we presume it's just doing these calculations. It's moving in matrices. It's working in some vector space. It's doing some higher dimensional thing. I have some experience of understanding. I guess people are still grasping at that. Is it having some experience of understanding?

Is it important whether or not they experience understanding? Is that sufficient to call it comprehension of meaning? Are you describing understanding as a behavioral trait here where it gives the right answers to problems or whether it deeply at the neural level understands? Yeah, I'm I'm completely at the whims of the philosophers here. No, I I don't know if I understand that at my at the human level, right? I can't tell you what process I'm executing at the moment either, right? But I do have some intuitive subjective experience that I understand the conversation. Obviously not that well. Um but but uh I when I'm talking to you, I feel you are understanding and when I'm talking to chat GPT, I do not. And you're telling me I'm mistaken.

It's understanding as well as I am or you are. In my opinion, it is understanding, yes. And I think there's two different pieces of evidence for that. One is I think if you talk to them like if you talk to them and ask them about difficult concepts I'm frequently surprised and with every passing month and every new model that comes out, I am more and more surprised at the level of sophistication with which they're able to discuss things.

And so just just at that level, it it's super impressive. I I would really encourage everybody here to talk to these large language models if you've not already. You know, when the science fiction writers imagined that we'd built some sort of Turing test passing machine that that was going to you know some new alien intelligence that we'd have in a box, they all imagined that we'd sort of hide it in a basement, you know, in a castle surrounded by a moat with armed guards and we'd only have like a priestly class who could be able to go and talk to it. That is not not it's not the way it worked out. The way it's worked out is the first thing we did is we immediately hooked it up to the internet and now anybody can go talk to it and I would highly encourage you to to talk to these things and explore in areas that you know to see both their limitations but also their strength and their their depth of understanding. So I'd say that's the first piece of evidence. The second piece of evidence is you said they're a black box. They're not exactly a black box. We do have access to their neurons. In fact, we have much better access to the neurons of these things than we do with a human.

It's very hard to get IRB approval to slice up a human while they're doing a math test and see how their neurons are firing and if you do do that, you can only do that once on on a per human basis. Whereas these neural networks, we can freeze them, replay them, write down everything that happened. If we're curious, we can go and go and prod their neurons in certain ways and see what happened. And so this is it's still rudimentary but this is the field of interpretability, mechanistic interpretability, trying to understand not just what they say but why they say it, how they think it.

And when you do that, we see uh when you feed them a math problem problem, there's a little bit of a a circuit there that computes the answer that that we didn't program it to have that. It learned how to do that. While trying to predict the next token on all of this text, it learned that in order to most accurately predict the next the next word, I should say. In order to most accurately predict the next word, it needed to figure out uh how to do

maths and it needed to build a sort of proto little circuit inside it to do the mathematical computations.

Mhm. Now, Ian, you famously threw a slide up at one of your uh keynote lectures that was very provocative, um very scholarly. It said um machine learning sucks, I believe was it. And then that kind of went wild. Ian LeCun says machine learning sucks. Um why are you saying machine learning sucks? Adam has just told us how phenomenal it is. He talks to them >> [laughter] >> and wants us to do the same. Um why do you think it sucks? What's the problem?

>> [clears throat] >> Well that statement has been widely misinterpreted but the point the point I was making is the point that uh we both we both made which is that why is it that a teenager can learn to drive a car in 20 hours of practice uh a 10-year-old can clean up the dinner table and fill up the dishwasher the first time you ask the child to do it. Whether the 10-year-old will want to do it is a different story, but you know, certainly can. Um we don't have robots that are anywhere near this.

And we don't have robots that are even anywhere near the you know, physical understanding of of reality of of a cat or a dog. And so in that sense, machine learning sucks. It doesn't mean that the the deep learning method, the back propagation algorithm, the neural nets suck. That was obviously excellent. Yes. >> Yes, obviously. That's great. And [laughter] we don't have any alternative to this. And uh I I certainly believe that you know, neural nets and deep learning and back propagation will be you know, are with us for for a long time. We'll be the basis of uh future AI systems. But but how is it that uh uh you know, young humans can can learn how the world works in the first few months of life?

It takes 9 months for a human babies to learn um intuitive physics like gravity, inertia, and things like this. Um baby animals learn this much faster. They have smaller brains, so it's easier for them to learn. Um They don't learn to the same level, but they but they do learn faster. And and so, you know, there is this type of learning that we need to reproduce. Um and we'll do this with back prop, with neural nets, with deep learning. It's just that we're missing a concept, an architecture.

Um so I've been I've I've been making proposals for the type of architectures that could possibly uh learn this kind of stuff. You know, why is it that LLMs handle language uh so easily? It's because um as Adam described, you you train an LLM to predict the next word or the next token, doesn't matter. There's only a finite number of words in the dictionary. So you can never actually predict exactly which word comes after a sequence.

But you can train a system to produce essentially what amounts to a score for every possible words in your dictionary or a probability distribution over every possible words. So essentially what an LLM does is that it produces a long list of numbers between zero and one that sum to one, which for each word in the dictionary says, this is the likelihood that this word appears right now. You can represent the uncertainty in the prediction this way. Now, try to translate it um the the same principle. Instead of training a system to predict the next word, um feed it with a video and ask it to predict what happened next in the video.

And this doesn't work. I've been trying to do this for 20 years. And it it really doesn't work if you try to predict

at a pixel level. Uh and it's because the real world is messy. There's a lot of things that that may happen, plausible things that may happen. Um and you can't really represent a distribution over all possible uh things that may happen in the future because it's basically an infinite list of possibilities and we don't know how to represent this um efficiently. And so those those techniques that work really well for text or for sequences of of symbols do not work for real world sensory data.

They just don't They absolutely don't. And and so we need to invent new techniques. So one of the things I've been proposing is one in which the the system learns an abstract representation of what it observes and it makes prediction in that abstract representation space. And this is really the way humans and animals function. And we we find abstractions that allow us to make predictions while ignoring all the detail the details we cannot predict. So you really think that despite the phenomenal successes of these LLMs that they are limited and and their limit is quickly approaching. You don't think that they're scalable to this you know, artificial general intelligence or super intelligence.

>> That's right. No, they don't. [clears throat] And and in fact, we see the performance saturating. So we we see uh progress in in some domains like mathematics, for example. And mathematics and and code generation, you know, programming are two domains where the uh the the manipulation of symbols actually gives you something, right? As a physicist, you you know this, right? You write the equation and it actually kind of You follow it. You you can follow it and it it's uh it drives your your thinking into some extent, right? I mean, you you drive it by intuition, but but the symbol manipulation itself actually has uh meaning. So this type of problems, LLMs actually can handle pretty well, where the the reasoning really consists in kind of searching through sequences of symbols. But it's only there's only a small number of problems for which that's the case.

Chess playing is another one. Um you search through sequences of of of moves that, you know, for a good one or sequences of uh derivations in mathematics that will produce a particular result, right? Um but in the real world, it you know, in high-dimensional continuous things where the search has to do with like how do I move my muscles to uh you know, grab this uh you know, grab grab this um this this glass here. I'm not going to do it with my left hand, right? I'm going to have to change hand with this and and then grab it, right? You need to plan and have some understanding of what's possible, what's not possible, that, you know, I can't just attract the glass, you know, by telekinesis or I can't just I can't just make it appear in my in my left hand like this. I can't have my hand kind of cross my body. Like you know, all of those intuitive things, we we learn them when we were babies. Um and and we learn, you know, how our body reacts to our controls and how uh you know, the the world reacts to to the actions we take. So so, you know, if I push this glass, I know it's going to slide. If I push it from the top, maybe it maybe it's going to flip. Maybe not because the friction is not that high. If I push with the same force on this table, it's not going to flip.

You know, we have all those those intuitions that allow us to kind of apprehend the real world. Uh but this is it turns out much much more complicated than manipulating language. We think of language as kind of the epitome of you know, human intelligence and stuff like that. It's actually not true. Language is actually easy. >> [laughter] >> Is it the Moravec paradox that what computers are good at, humans are bad at? What humans are good at, computers are bad at? Yeah, we keep running into the Moravec paradox here. Now, Adam, I I know

that you are less pessimistic about the potential of the current neural net deep learning um paradigm and you see the potential for a great escalation in success and uh you don't see it saturating. Um what's your thought about that?

>> I am I don't. That's right. Um And so yeah. >> witnessed over the last 5 years the most extraordinary run-up in capabilities in any system I've ever seen. This is what transfixed my attention. It's what transfixed many other people uh in AI and neighboring fields to focus all of our attention on this matter.

I don't see any slowdown in the capabilities. A year ago if you just look at all of the all of the metrics we used to judge how good these large language models are, they're getting stronger and stronger and stronger. Things that they, you know, a year the model from a year ago today would be, you know, table stakes, would be considered extremely poor. Every few months these things push their capabilities. And if if you track their capabilities on all of these tasks, they're heading towards superhuman on on almost all of them. It's already better gives better legal advice than uh than a lawyer. It gives better um it's a better poet than almost every poet you will come. In my little in my little area of physics, uh I I use it because like there's something I kind of should know, but I don't. I'll ask the language model and it will not only tell me what the right answer is, it will patiently and I should say non-judgmentally listen while I explain my misconception to it and it will carefully debunk my misconception.

Um the extraordinary run-up in capabilities that we've seen over the last 5 years uh and it continues up to the present is extremely tantalizing to to me and many other people in San Francisco. And and maybe maybe Jan is correct that we're just going to suddenly saturate and all of these uh straight lines that have been going up steadily for the last 5 years are suddenly going to stop going up, but I am mighty curious to see uh whether we can push it further. And I've actually seen no indication whatsoever that it's slowing down. Every indication I've seen is that these these are improving. And we don't have far to go because once it's a better coder than almost all our best coders, it can start improving itself. And then we're really in for a wild ride.

Well, we we've had better coders than the original coders of 1950s, >> [clears throat] >> you know, for six decades or so. That's called compilers. I mean, we we we keep getting confused about the fact that it's not because machines are good at a certain number of tasks that they have all the underlying intelligence that we assume a human having those capabilities will have, right? We're fooled into thinking those machines are intelligent because they can manipulate language. And we're used to the fact that people who can manipulate language very well are implicitly smart. Um but we're being fooled. Um now, they're useful, there's no question. Um you know, we can use them to do what you said. I I them for similar things.

Great. They're great tools like, you know, computers uh have been for the last few decades five decades. But, let me make an interesting historical point. >> [snorts] >> Um and this is maybe due to my age. Uh there's been generation after generation of AI scientists since the 1950s claiming that the technique that they just discovered was going to be the ticket for human-level intelligence. You you see declarations of Marvin Minsky, Newell and Simon, um you know, uh Frank Rosenblatt who invented the perceptron, the first learning machine in 1950 saying like within 10 years we'll have machines that are as smart as humans. They were all wrong. This

generation with LLM is also wrong. I've seen three of those generation in my lifetime, okay?

Um so, you know, it's it it's just another example of being fooled. And um in the '50s, Newell and Simon, pioneers of AI, came up with a program that said, "Well, you know, really what what humans are doing um is in reasoning is really a search, right? Every reasoning can be reduced to kind of a kind of search. So, you formulate a problem, you write a program that tell you whether a particular proposal for a solution is a solution to your problem, and then you just have to search for all possible combinations, you know, all possible hypotheses for one that actually matches uh satisfies the the constraint. And that's it. We're going to write a program that does this, and we're going to call it the general problem solver, GPS, 1950 seven, I think.

Uh they won the Turing Award for for things like that. And it was it was great, but then they didn't realize that all the interesting problems actually have a complexity that goes exponentially with the size of the problem. So, in fact, you can't really use this uh uh technique to build uh intelligent machines. It can be a component of it, but it's really not not the thing. So, simultaneously uh Frank Rosenblatt came up with a perceptron, a machine that could learn.

He said, "If you can train a machine, then it can become infinitely smart." And so, within 10 years we'll have we just need to big, you know, to build bigger perceptrons, right? Not realizing that you need to train multiple layers, and that turned out to be uh difficult to find a solution for this. Um then in the 1980s there was uh expert systems, okay? Reasoning is is fine. Just write a bunch of facts and a bunch of rules, and then just deduce all the facts from the original facts and the and the rules, and uh now we can reduce all the human knowledge into into this.

The the coolest job is going to be knowledge engineer. You're going to sit down next to an expert, and then write down all the rules and the facts, and turn this into an expert system. And, you know, everybody was excited about this, and there were, you know, billions that were invested. The Japan started the fifth generation computer program pro project, which was which was going to revolutionize computer science. Complete failure. Okay? It created an industry. It was useful for a few things, but basically the cost of reducing human knowledge to to rules uh was just too high for most problems, and so the whole thing collapsed. Then there was neural nets, the the first the second wave of neural nets 1980s deep, you know, which we now call deep learning.

A lot of interest, but then it was before the internet. We didn't have enough data. We didn't have powerful computers. And now we're we're going through the same cycle again, and we're getting fooled again. So, just to be Oh, Adam, please. In in technologies every dawn has before it false dawns. That doesn't mean we'll never we'll never hit the dawn. I I guess I would like um Yann, if you think that LLMs are going to saturate, what is a concrete task that they will never be able to do?

That a that an LLM augmented by, you know, the the tools we give it today will never be able to perform. Uh Uh clear the dinner table, fill up the dishwasher. >> [laughter] >> Okay. Okay. And that's easy compared to I'm skeptical. That's super easy compared to fixing your toilets. Yeah. Okay? Plumber, right? You're never going to have a plumber with LLMs. You're never going to have a robot driven by LLMs. It just cannot understand the

real world. It just can't. So, I want to clarify for the audience that you're not saying that machines or robots won't be able to do this. That's not your position. You think they will.

>> They will. They absolutely will. >> Just not by this algorithmic approach or for this particular approach of the deep learning on the neural nets. >> program I'm working on succeeds, which may take a while, >> Jeppa? Am I Jeppa? Jeppa? Jeppa and and, you know, all the things world models and things that go with it. If it succeeds, which may take, you know, several years, then we we may have, you know, AI system There's no question that at some point in the future we will have machines that are smarter than humans in all domains that you know, where humans have abilities. There's no question about that. It will happen, okay? It probably take longer than, you know, some of the people in Silicon Valley at the moment are seeing it it it will. Uh and uh and it it will not be LLMs. It will not be generative models that predict discrete tokens.

It will be models that learn abstract representations and make predictions on abstract representations and can reason about what is going to be the effect of me taking this action, can I plan a sequence of actions to arrive at a particular goal. You call this self-supervised learning. No, so self-supervised learning is used also by LLMs. Self-supervised learning is the idea that you train a system not for a a particular task other than capturing the the sort of underlying structure of the of the data you you you show it. And one way to do this is to give it a piece of of data, corrupt it in some way by uh removing a piece of it, for example, masking a piece of it, and then training a bit neural net to predict the piece that is missing.

So, LLMs do this, right? You take a text, you remove the last word, and you train the LLM to predict the the word that is missing. You have other types of language models that actually fill up multiple words. They turned out to not work as well as the ones that just predict the last one. Um at least for certain task. Um you can do this with video. If you try to predict at the pixel level, it doesn't work or it doesn't work very well.

Um My colleagues at Meta probably boiled a couple small lakes in the West Coast to you know, trying to make this work. Um >> [laughter] >> to cool the GPUs. Uh so, it simply doesn't work. So, so you have to, you know, come up with those new architectures like Jeppa and stuff like that. And those kind of work. Like we we have models that actually understand video. And Adam, are people exploring other ways of building an architecture or imagining a computer mind this the actual fundamental structure of a computer mind and how it would um how it would learn, how it would acquire information.

One of the criticisms, as I understand it, is it's a lot of the LLMs are trained for this one specific task of this discrete prediction of these um tokens. But, something that is more unpredictable like how the audience is distributed in this room, what might happen with the weather next to unpredictable more human experience-based phenomena. Um certainly all kinds of explorations are being made in all kinds of directions uh including Yann's, and you know, let a thousand flowers bloom. Um But, all of the resources I mean, the the bulk of the resources right now are going into large language models and large language model-like applications in in including taking in text.

To say to say that they are it's a specialized task predicting the next token. I think that's a not a helpful way to

think about it. It is true that the thing that you train them on is given this corpus of text. I mean, there are other things we do as well, but the the bulk of the compute goes to given this corpus of text, please predict the next word. Please predict the next word. Please predict the next word.

Um but, we have discovered something truly extraordinary by doing it, which is that given a large enough body of text to be able to reliably predict the next word or, you know, do it do it well enough to predict the next word, you really need to understand the universe. And we have seen the emergence of understanding of the universe as we've done that. So, I I would liken it a a little bit. I mean, in physics we're very used to systems where you just take a very simple rule, and you know, by the repeated application of that very simple rule, you get extremely impressive behavior. Uh we see the same with these LLMs.

Uh and another example of that would maybe be evolution. You know, at each stage in evolution you just say uh biological evolution you just say, you know, maximize the number of offspring. Maximize the number of offspring. Maximize the number of offspring. A very sort of unsophisticated learning objective. But, out of this simple learning objective repeated many, many times, uh you eventually get all of the, you know, splendor of biology that we see around us and and indeed this room. So, the evidence is that predicting the next token, while a very simple task, because it's so simple we can do it at massive scale huge amounts of compute. And once you do it at huge amounts of compute, you get an emergent complexity.

Mhm. [clears throat] So, I I guess the next question could be related to evolution. However, this intelligence emerges that you both imagine is certainly possible. You don't think there's anything special about this wetware. That there will be machines we just have to figure out how to launch them that will um have capacities that we align as a kind of intelligence or maybe consciousness. That's a almost a different question. Will consciousness be a crutch machines don't need? I don't know. We can talk about that. But, but is there a point in the evolution of these, uh machines where they're going to say, oh how quaint mom and dad, you you made me in your image with these human neural nets, but I know a way, a much better way having scanned 10,000 years of human output to make machine intelligence and I'm going to evolve and leave us in the dust. I mean, yeah, what do we why are we imagining that they would be limited at that capacity to the way we design them?

Uh absolutely. This is this idea of recursive self-improvement where when they're bad that they're useless, but when they get good enough and strong enough, you can start using them to augment human intelligence and perhaps eventually just be fully autonomous and replace and make future versions of them. Once we do that, I mean I think what we should do is just take this large language model paradigm that's currently working so well and just see how far we can push it. You know, it keeps every time someone says there's a barrier, it pushes through the barrier of the last 30 years, but eventually these things will get smart enough and then they can read Yann's papers, read all the other papers that have been made, try and figure out new ideas that none of us thought of.

Yeah. So, I completely disagree with this. Um So, LLMs are not controllable. It's not dangerous because they're not that smart as as I explained previously. Uh and they're certainly not autonomous in a way that we understand autonomy. We have to distinguish between autonomy and intelligence. You can be very intelligent

without being autonomous. And you can be autonomous without without being intelligent. Um and you can be dangerous without being particularly intelligent.

Um And you can want to be dominant without being intelligent. In fact, that's going to be inversely correlated in the human species. >> [laughter] >> Um >> [laughter] >> You know, politics. Um I won't cite names. So, I think what what is required is systems that are intelligent, in other words, can solve problems for us.

But they will solve the problem we give them. Okay? And again, that would require a new design than LLMs. LLMs are not designed to fulfill a goal. They're designed to predict the next word. And we fine-tune them so that they behave, you know, for particular questions, they answer in a particular way. Um but there's always what's called a generalization gap, which means you can never train them for every possible question. And there's a very long tail.

And so, they're not controllable. Um And [snorts] again, that doesn't mean they're very it's very dangerous because they're not that smart. Um Now, if you build systems that are smart, we want them to be controllable and we want them to be driven by objectives. We give them an objective and the only thing they can do is fulfill this objective according to their, you know, internal model of the world if you want. So, plan a sequence of actions that will fulfill that objective. If we design them this way and we also put guardrails in them so that in the process of fulfilling the objective, they don't do anything, you know, bad for for humans.

So, the the usual joke is if you have a robot, um domestic robot and you ask it to fetch you coffee and someone else is, you know, someone is standing in front of the coffee machine, you don't want your robot to just, you know, kill that person to get access to the coffee machine, right? So, you want to put some guardrail into the the behavior of that robot and we do have those guardrails in our head.

Evolution built them into us. Right? So, we don't kill each other all the time. I mean, we do kill each other all the time, but not, you know, not all the time all the time. Um I mean, you know, and you know, we feel empathy and things like that and that's just built into us by evolution. That that's the way evolution sort of hardwired guardrails into us. So, we should build our AI systems the same way. Have objectives and goals, drives, but also um you know, guardrails, inhibition, basically. Um And and then they will solve problems for us. They will amplify our intelligence. They will do what we ask them to do.

And our relationship to those intelligent system will be like the relationship of, let's say, a professor with graduate students who are smarter than them, right? >> Right. >> [laughter] >> I mean, I don't know about you, but I have students who are smarter than me. So, um It's the best thing that can happen to you, right? >> It's the best thing that can happen. Right? So, we'll be walking around with AI assistant um that will help us in our daily lives. They'll be smarter than us, but they'll work for us. They'll be like our staff.

Again, there is a political analogy here, right? A politician, right? He's a figurehead and they have a staff of people, all of all of whom are smarter than them, right? Um so, it's going to be the same thing with AI system, which is why I to the question of Renaissance, I said Renaissance. So, you have no concerns um about the safety

of the current models, but the question is maybe we should stop there. I mean, why is it necessary for us to scale up so widely that every single person has uh this super intelligence in their pocket on their iPhone. Is that really necessary?

A friend of mine was saying it's like bringing a ballistic missile [clears throat] to a knife fight. I mean, is this necessary that every person has a ballistic missile capability um or should we stop here where we have these controllable systems? >> You can say exactly the same thing about teaching people to read, giving the giving them a textbook of chemistry of volatile chemicals, you know, with which they can make explosives or nuclear physics book, right? I mean, we do not question the idea that knowledge and more intelligence is good, intrinsically good, right? We do not question anymore the fact that the invention of printing press was a good thing, right? It made everybody smarter.

It it gave it gave access to knowledge to everyone. Um Which was not possible before. It incited people to learn to read. It's it caused the enlightenment. It also caused 200 years of, you know, religious wars in Europe, but okay, but We got over it. Yeah. But it it caused the enlightenment. It caused, you know, the emergence of philosophy, science, democracy, the American Revolution, the French Revolution. All of that would not have been possible without the the printing press. So, you know, every technology that particularly communication technology, but technology that amplifies human intelligence, I think is intrinsically good. Now, Adam, people are concerned. I'm sure they'll feel very reassured.

That Yann is not concerned when these doomsday scenarios you think are highly exaggerated, but are you concerned about some of the safety issues around AI or our ability to really uh keep the relationship balanced in the direction that we want it to be? Um I think to the extent that I think this is going to be a more powerful technology than Yann thinks it does, I am more concerned. I think it's going to be a very powerful to the extent that it is a very powerful technology, it'll have both positive and negative impacts.

Um and I think it's very important to make sure that, you know, that we work together to make sure that the positive impacts are uh outweigh the negative impacts. I think that path is totally open to us. There are a huge number of possible positive impacts and we could just, you know, talk about some of those perhaps, but uh we need to make sure that that happens. Now, let's talk about agentic misalignment, which is the phrase that's been passed along. It was my understanding there was reports recently that when Claude 4 was rolled out, that those in simulations and tests, one of the models was or I don't know if there's a singular model. I don't know if it thinks it's of itself as a singular entity or they, but the model uh exhibited resistance to rumors in the simulation that it was going to be replaced. It was sending messages to its future self um trying to undermine the intentions of the developers. It faked legal documents and it threatened to blackmail one of the engineers. Right? [laughter] Um so, this notion they were concerned.

Um uh so, this notion of agentic misalignment, is that something that you're concerned with that there will be a power over, say, financial systems, heating and cooling systems, the energy grid, and um and that that they will resist its developers' intentions. Yeah, so that that paper was a paper by Anthropic, which is a paper in company in San Francisco, not my company, but a company that takes safety very seriously. And they did a slightly mean

thing to their LLM where they gave it a scenario, sort of philosophy professor style scenario where it had to do a bad thing to stop an even worse thing happening. Sort of, you know, utilitarian ethics and deontological ethics colliding and it was eventually persuaded by them to do the utilitarian thing. And that's kind of not what we we want, I would say. We want it that if it has a rule that, you know, will not lie, that it will not lie, no no matter what.

Um and to their credit, they tested it for that, found that it would occasionally act deceptively if if promised that by doing so it could save that many lives. These are tricky things that, you know, human philosophers wrestle with. Um I think it is a We need to be careful to train them to obey our command. And and and we spend a lot of time doing that. >> Who's us? Um isn't there's a big concern uh we're assuming that all of humanity is aligned in our intentions. That's clearly not the case. And And I know you and you in a very interesting way argue for open source, which some people would say is even more dangerous cuz now anyone can have access to it. It's dangerous enough that it's in the hands of a a a small number of people who rule corporations, but let alone everyone having it. Maybe that is dangerous. Um But again, who's us and we? The danger is if we don't have open source AI systems.

Okay, in the future, every single one of our interaction with a digital world will be mediated by an AI system. Right? We're not going to go to a website or search engine or whatever. We're just going to talk to our AI assistant. However, it's built. Um so our entire information diet will come from AI assistants. Now, what does it mean to culture, language, democracy, uh everything if those systems come from a handful of companies on the West Coast of the US or China?

I tell you, no country in the world outside the US and China likes the idea. Um so we need a high diversity of AI assistant for the same reason we need a high diversity of the press. We cannot afford to have just a handful of proprietary system coming out of a small number of companies. There's one thing I'm scared of, and that's it. Okay? If we don't have open platforms, um we're going to have uh you know, capture of information flow by a handful of companies, some of which we may not like.

And so So, how can we be certain that these um when when they really are self-motivated agents, if that ever actually happens, that they won't collude, fight amongst themselves, want to wrestle for power, that we won't be sitting back watching conflicts that we simply couldn't have imagined before? We give them clear objectives, and we build them in such a way that the only thing they can do is fulfill those objectives. Now, this is not doesn't mean it's going to be perfect, but the question of AI safety in the future, I'm I'm worried about it in the same way that I'm worried about the question of reliability of turbojets. Okay?

I mean, turbojets, I mean, it's it's amazing. I don't know about you, but and my dad was aeronautical engineer, but I'm totally amazed by the fact that you can fly halfway around the world in complete safety on a two-engine airplane. It's amazing, right? And and we feel completely safe doing this. It's a It's It's a magical production of uh you know, engineering of the modern science and engineering of the modern world. AI safety is a problem of this type.

It's It's an engineering problem. Um I think the fears are caused by people who think about um you know, science fiction scenario where somewhere someone invents the secret to superintelligence, turns on the machine, and the next second it takes over the world. That is complete BS. Like, the world doesn't work this way. Certainly the world of technology and science doesn't work world this way. The emergence of superintelligence is not going to be an event.

Um as we see, we have superintelligent systems that can do superintelligent tasks, you know, and there is kind of continuous progress one at a time. Uh but you know, we're going to find some, you know, better recipe to build AI systems that may have kind of a more general intelligence than we currently have. Um we'll have systems There's no question that are smarter than humans. But we'll build them so that they fulfill the goals we give them, subject to guardrails.

Um I I I was going to uh again question this idea of we we we We know that if we can code them in a certain way, somebody could recode them. And the concept of bad actors, but before we fall into that hole, I have a plant in the audience. Does my plant have a mic? Is my plant know who he is? Does my Meredith Isaac, does my plant have a mic? Yes? He's up there. Oh, but he doesn't have the mic. Okay, David, can you shout?

No. Okay, so >> [laughter] >> um so I want to introduce the uh philosopher of mind, David Chalmers. I'm going to give you a very brief introduction. David, I can't see you, but I I I said um that you could be my plant to ask a question. Could you Do you want to throw something down here? Okay, I'm down here. Okay, you asked Janet asked me to ask a question about uh AI and consciousness. Hi, Adam. Hi, Janet.

OKAY, SO UH YOU BOTH SAID, I THINK, ROUGHLY, current AI systems probably not conscious. Future AI systems, possibly descendants of the ones today, but some future AI systems probably will be conscious. So I guess I want to know uh one, what requirements for consciousness do you think current systems are lacking? Um and then the positive side of that is um what steps do you think we need to take in order to develop AI systems which are conscious?

And then third, when is that going to happen? Okay, I give a card at this. Uh indeed, he already knows my answer, but um So first of all, I don't attribute like I don't really know how to define consciousness, and I don't attribute much importance to it. And this is an insult to David. I'm sorry. Uh because he devoted his entire career to it. >> Subjective experience. Okay, that's a different thing. Okay, subjective experience.

Um So clearly, we're going to have systems that have subjective experience, that have emotions. Emotions, to some extent, are an anticipation of outcome. If we have systems that have world models that are capable of anticipating uh the outcome of a situation, perhaps resulting from their actions, they're going to have emotions because they can predict whether something is going to end up, you know, good or bad for you know, in on the way to fulfilling their objectives, right? So So they're going to have all those characteristics. Now, I don't know how to define consciousness in this kind of in in this, but perhaps uh consciousness would be the ability for the system to kind of observe itself and configure itself to solve a particular subproblem that it's facing.

It needs to have kind of a way of observing itself and configuring itself to um solve a particular problem. We We certainly can can do this. And so um perhaps that's what gives us the illusion of of consciousness. I don't I haven't I have no doubt this will happen at some point. >> And will the machines have moral worth when it happens? Yeah, absolutely. I mean, they will have some moral sense. Whether it aligns with us or not will depend on how we define those objectives and guardrails.

Um but yeah, they will have a sense of of moral. Let me ask Adam this question a slightly different way, or you can answer the same question as well. Um Are we too attached to the human subjective experience, our sense of consciousness? Uh clearly, we've already know that animals don't have the same experience that we do. And uh why should we imagine that the superintelligence will have the same subjective experience as human beings? Okay, let me answer all your questions, then. Uh just my my gut. I I think machines can certainly be conscious in in in principle, that if they're doing at the you know, the artificial neurons and they're doing the same information processing in the same way as human neurons, uh then then you know, the very least that will give rise to to consciousness. It's not about the substrate, whether it's silicon or carbon. It's just about the nature of the information processing will give rise to consciousness.

Um what we're missing to get there, um as as David knows, there's, you know, there are these things called the neural correlates of consciousness. People who don't want to say they're studying consciousness directly could look at human brains, or perhaps animal brains, and say, "What is the processes going on in the neurons that give rise to conscious experience?" Um and uh there's a number of number of theories, and from my point of view, they all kind of suck. Um There's There's the recurrence theory that you need to be able to take your outputs and plug them back in to the inputs, and that's an essential part of consciousness. There's something called global workspace theory, integrated information theory. Every, you know, physicist and neuroscientist like to have their own def- set of criteria for what it is for a machine for a information processing system to be conscious. I don't find any of them particularly compelling, and I think we should have extreme humility about recognizing consciousness in other entities. We are very bad at doing it in, you know, in animals. We very much changed our mind over history whether animals are conscious, uh whether babies experience consciousness. So, my question is a little bit don't know.

But, I do think that if you just told me about neural networks or told, you know, if I if I didn't know about consciousness and I just hear heard about the processing of information that happens in neural neural networks, human neural networks, I would not have predicted that gives rise to consciousness. That's a great surprise. And we should be for that reason extremely humble even about what the form of the consciousness would make. So, to to answer Janet's question, we have seen that what we used to think of as a reasonably unified idea of intelligence, human intelligence, which is a whole bunch of different abilities and and skills, we've unbundled that with these machine intelligences where we constructed things that have some of them but not others. Very superhuman in some, subhuman in others. Perhaps we will be unbundling consciousness as well and this thing that we think of as consciousness, we will realize that there is many different aspects to it that we can have some and not the others and maybe as you indicated we can even transcend human consciousness in in some capacities. I'm pretty excited about answering this question though. I I think we finally finally finally have a model organism for intelligence in the form of these artificial minds that we're building and maybe we can talk turn that model organism for intelligence into a model organism for consciousness and answer some of these

questions that have intrigued mankind.

I just didn't think I heard an answer to when. Oh, [laughter] um I I can neither confirm nor deny I think is the standard phrase we're using here. >> [laughter] >> Um I think if if progress keeps going, uh 2036. >> [laughter] >> Okay, not in the next 2 years. Um just one closing question. We're a little bit over time, but let me ask this to you, Jan. Uh in many ways you're a contrarian, maybe not by choice. Maybe this is just how it's happened. You've called it the cult of LLMs. You you sort of often refer to the fact that in Silicon Valley you don't have the most conventional approach. Um but yet you have an optimism.

>> [clears throat] >> You you really do not indulge in the doomsday sort of rhetoric. Uh what is your most optimistic vision for if not 2 years from now, 2036? Well, the new Renaissance. That's a pretty optimistic uh view of, you know, AI systems that amplify human intelligence is under control, can solve uh a lot of complex problems, can accelerate the progress of science and medicine, can uh educate our children, um you know, help us uh you know, process all the information or bring us uh all the knowledge and information that we need to to see.

Uh in fact, you know, people have been interacting with AI systems for much longer than they realize. Um of course, there is, you know, LLMs and chatbots now for the last 3 years. Uh but before that, um you know, most like every car sold in in in the EU and most cars sold in in the US have uh what's called ADAS, advanced driving assistance systems, or automatic emergency braking systems. You know, a camera that looks out the window and stops your car if you're about to hit a pedestrian or another car. Um it saves lives. Um you you get a an X-ray today, let's say a a mammogram or something. You know, at the bottom it says the thing has been reviewed by an AI system. It saves lives. Um you can get an MRI now, full body MRI in 40 minutes.

Um this is because you can accelerate the process of collecting the data because AI systems can uh can sort of fill in the blanks. You don't need to collect as much data for this. Um but also all the news you're seeing, whether you connect on Google or, you know, Facebook, Instagram, any social network, is determined by an AI system that basically caters to your interests. Um and so, you know, AI has been with us for for a while already.

But you're saying we should be impressed when they can pour a glass of water and do our dishes. >> of water, do our dishes, uh you know, drive our cars, like learn to drive our cars in 10 hours without the cheating. Yes. without all the the cheating with with sensors and mapping and and and and hardcoding of rules. So, um yeah, this is going to take a while. Uh but this is going to be the next revolution of AI. So, this is what I'm working on, okay? Um and the the message I've been, you know, carrying for a while now is um is okay, LLMs are great, they're useful, we should invest in them. Um a lot of people are going to use them.

They are not a path to human level intelligence. They're just not. Uh right now they're sucking the air out of the room anywhere they go. And so, there's basically no resource left for anything else. Um and so, for the next revolution, we need to kind of you know, take a step back and figure out what's missing from um current approaches and then I've been making proposal on this and working inside of uh Meta for a number of years on this alternative approach. It's uh come to the point where um you know, we need to kind of accelerate this this

progress now because we know it works. We have early results and so, um that's the plan. Hm. Okay, I could we could have a whole 'nother hour starting right here, but um I hope you'll all join me in thanking our guests for an incredible conversation. Thank you [music] so much.

>> [applause]