Dwarkesh Podcast #90 - Satya Nadella – Microsoft's AGI Plan & Quantum Breakthrough

Published - February 19, 2025

Transcribed by - thepodtranscripts.com

Dwarkesh Patel
Satya, thank you so much for coming on the podcast.

In a second, we're going to get to the two breakthroughs that Microsoft has just made, and congratulations, same day in Nature: the Majorana zero chip, which we have in front of us right here, and also the world human action models. But can we just continue the conversation we were having a second ago? You're describing the ways in which the things you were seeing in the 80s and 90s, you're seeing them happen again.

Satya Nadella
The thing that is exciting for me... Dwarkesh, first of all, it's fantastic to be on your podcast. I'm a big listener, and I love the way that you do these interviews and the broad topics that you explore.

The thing that is exciting for me... It reminds me a little bit of my, I'd say, first few years even in the tech industry, starting in the 90s, where there was real debate about whether it's going to be RISC or CISC, or, "Hey, are we really going to be able to build servers using x86?"

When I joined Microsoft, that was the beginning of what was Windows NT. So, everything from the core silicon platform to the operating system to the app tier- that full stack approach- the entire thing is being litigated.

You could say cloud did a bunch of that, and obviously distributed computing and cloud did change client-server. The web changed massively. But this does feel a little more like maybe more full-stack than even the past that at least I've been involved in.

Dwarkesh Patel
When you think about which decisions ended up being the long-term winners in the 80s and 90s, and which ones didn't, and especially when you think about- you were at Sun Microsystems, they had an interesting experience with the 90s dotcom bubble. People talk about this data center build-out as being a bubble, but at the same time, we have the Internet today as a result of what was built out then.

What are the lessons about what will stand the test of time? What is an inherent secular trend? What is just ephemeral?

Satya Nadella
If I go back, the four big transformations that I've been part of, the client and the client-server. So that's the birth of the graphical user interface and the x86 architecture, basically allowing us to build servers.

It was very clear to me. I remember going to what is PDC in '91, in fact I was at Sun at that time. In '91, I went to Moscone. That's when Microsoft first described the Win32 interface and it was pretty clear to me what was going to happen, where the server was also going to be an x86 thing. When you have the scale advantages accruing to something, that's the secular bet you have to place. What happened in the client was going to happen on the server side, and then you were able to actually build client-server applications. So, the app model became clear.

Then the web was the big thing for us, which we had to deal with in starting, in fact as soon as I joined Microsoft, the Netscape browser or the Mosaic browser came out what, I think, December or November of '93, right? I think is when Andreessen and crew had that.

So that was a big game-changer, in an interesting way, just as we were getting going on what was the client-server wave, and it was clear that we were going to win it as well. We had the browser moment, and so we had to adjust. And we did a pretty good job of adjusting to it because the browser was a new app model.

We were able to embrace it with everything we did, whether it was HTML in Word or building a new thing called the browser ourselves and competing for it, and then building a web server on our server stack and go after it. Except, of course, we missed what turned out to be the biggest business model on the web, because we all assumed the web is all about being distributed, who would have thought that search would be the biggest winner in organizing the web? And so that's where we obviously didn't see it, and Google saw it and executed super well.

So that's one lesson learned for me: you have to not only get the tech trend right, you also have to get where the value is going to be created with that trend. These business model shifts are probably tougher than even the tech trend changes.

Dwarkesh Patel
Where is the value going to be created in AI?

Satya Nadella
That's a great one. So I think there are two places where I can say with some confidence. One is the hyperscalers that do well, because the fundamental thing is if you sort of go back to even how Sam and others describe it, if intelligence is log of compute, whoever can do lots of compute is a big winner.

The other interesting thing is, if you look at underneath even any AI workload, like take ChatGPT, it's not like everybody's excited about what's happening on the GPU side, it's great.

In fact, I think of my fleet even as a ratio of the AI accelerator to storage, to compute. And at scale, you've got to grow it.

Dwarkesh Patel
Yeah.

Satya Nadella
And so, that infrastructure need for the world is just going to be exponentially growing.

Dwarkesh Patel
Right.

Satya Nadella
So in fact it's manna from heaven to have these AI workloads because guess what? They're more hungry for more compute, not just for training, but we now know, for test time. When you think of an AI agent, it turns out the AI agent is going to exponentially increase compute usage because you're not even bound by just one human invoking a program. It's one human invoking programs that invoke lots more programs. That's going to create massive, massive demand and scale for compute infrastructure. So our hyperscale business, Azure business, and other hyperscalers, I think that's a big thing.

Then after that, it becomes a little fuzzy. You could say, hey, there is a winner-take-all model- I just don't see it. This, by the way, is the other thing I've learned: being very good at understanding what are winner-take-all markets and what are not winner-take-all markets is, in some sense, everything. I remember even in the early days when I was getting into Azure, Amazon had a very significant lead and people would come to me, and investors would come to me, and say, "Oh, it's game over. You'll never make it. Amazon, it's winner-take-all."

Having competed against Oracle and IBM in client-server, I knew that the buyers will not tolerate winner-take-all. Structurally, hyperscale will never be a winner-take-all because buyers are smart.

Consumer markets sometimes can be winner-take-all, but anything where the buyer is a corporation, an enterprise, an IT department, they will want multiple suppliers. And so you got to be one of the multiple suppliers.

That, I think, is what will happen even on the model side. There will be open-source. There will be a governor. Just like on Windows, one of the big lessons learned for me was, if you have a closed-source operating system, there will be a complement to it, which will be open source.

And so to some degree that's a real check on what happens. I think in models there is one dimension of, maybe there will be a few closed source, but there will definitely be an open source alternative, and the open-source alternative will actually make sure that the closed-source, winner-take-all is mitigated.

That's my feeling on the model side. And by the way, let's not discount if this thing is really as powerful as people make it out to be, the state is not going to sit around and wait for private companies to go around and... all over the world. So, I don't see it as a winner-take-all.

Then above that, I think it's going to be the same old stuff, which is in consumer, in some categories, there may be some winner-take-all network effect. After all, ChatGPT is a great example.

It's an at-scale consumer property that has already got real escape velocity. I go to the App Store, and I see it's always there in the top five, and I say "wow, that's pretty unbelievable".

So they were able to use that early advantage and parlay that into an app advantage. In consumer, that could happen. In the enterprise again, I think there will be, by category, different winners. That's sort of at least how I analyze it.

Dwarkesh Patel
I have so many follow-up questions. We have to get to quantum in just a second, but on the idea that maybe the models get commoditized: maybe somebody could have made a similar argument a couple of decades ago about the cloud – that fundamentally, it's just a chip and a box.

But in the end, of course, you and many others figured out how to get amazing profit margins in the cloud. You figured out ways to get economies of scale and add other value. Fundamentally, even forgetting the jargon, if you've got AGI and it's helping you make better AIs – right now, it's synthetic data and RL; maybe in the future, it's an automated AI researcher – that seems like a good way to entrench your advantage there. I'm curious what you make of that, just the idea that it really matters to be ahead there.

Satya Nadella
At scale, nothing is commodity. To your point about cloud, everybody would say, "Oh, cloud's a commodity." Except, when you scale... That's why the know-how of running a hyperscaler... You could say, "Oh, what the heck? I can just rack and stack servers."

Dwarkesh Patel
Right.

Satya Nadella
In fact, in the early days of hyperscale, most people thought “there are all these hosters, and those are not great businesses. Will there be anything? Is there a business even in hyperscale?” And it turns out there is a real business, just because of the know-how of running, in the case of Azure, the world's computing of 60-plus regions with all the compute. It's just a tough thing to duplicate.

So I was more making the point, is it one winner? Is it a winner-take-all or not? Because that you've got to get right. I like to enter categories which are big TAMs, where you don't have to have the risk of it all being winner-take-all. The best news to be in is a big market that can accommodate a couple of winners, and you're one of them.

That's what I meant by the hyperscale layer. In the model layer, one is models need ultimately to run on some hyperscale compute. So that nexus, I feel, is going to be there forever. It's not just the model; the model needs state, that means it needs storage, and it needs regular compute for running these agents and the agent environments.

And so that's how I think about why the limit of one person running away with one model and building it all may not happen.

Dwarkesh Patel
On the hyperscaler side, and by the way, it's also interesting the advantage you as a hyperscaler would have in the sense that, especially with inference time scaling and if that's involved in training future models, you can amortize your data centers and GPUs, not only for the training, but then use them again for inference.

I'm curious what kind of hyperscaler you consider Microsoft and Azure to be. Is it on the pre-training side? Is it on providing the O3-type inference? Or are you just, we’re going to host and deploy any single model that's out there in the market, and we are sort of agnostic about that?

Satya Nadella
It’s a good point. The way we want to build out the fleet is [to], in some sense ride Moore's law. I think this will be like what we've done with everything else in the past: every year keep refreshing the fleet, you depreciate it over whatever the lifetime value of these things are, and then get very very good at the placement of the fleet such that you can run different jobs at it with high utilization. Sometimes there are very big training jobs that need to have highly concentrated peak flops that are provisioned to it that also need to cohere. That's great. We should have enough data center footprint to be able to give that.

But at the end of the day, these are all becoming so big, even in terms of if you take pre-training scale, if it needs to keep going, even pre-training scale at some point has to cross data center boundaries. It's all more or less there.

So, great, once you start crossing pre-training data center boundaries, is it that different than anything else? The way I think about it is hey, distributed computing will remain distributed, so go build out your fleet such that it's ready for large training jobs, it's ready for test-time compute, it's ready- in fact, if this RL thing that might happens, you build one large model, and then after that, there's tons of RL going on. To me, it's kind of like more training flops, because you want to create these highly specialized, distilled models for different tasks.

So you want that fleet, and then the serving needs. At the end of the day, speed of light is speed of light, so you can't have one data center in Texas and say, "I'm going to serve the world from there."

You've got to serve the world based on having an inference fleet everywhere in the world. That's how I think of our build-out of a true hyperscale fleet.

Oh, and by the way, I want my storage and compute also close to all of these things, because it's not just AI accelerators that are stateless. My training data itself needs storage, and then I want to be able to multiplex multiple training jobs, I want to be able to then have memory, I want to be able to have these environments in which these agents can go execute programs. That's kind of how I think about it.

Dwarkesh Patel
You recently reported that your yearly revenue from AI is $13 billion. But if you look at your year-on-year growth on that, in like four years, it'll be 10x that. You'll have $130 billion in revenue from AI, if the trend continues. If it does, what do you anticipate doing with all that intelligence, this industrial scale use?

Is it going to be through Office? Is it going to be you deploying it for others to host? You've got to have the AGIs to have $130 billion in revenue? What does it look like?

Satya Nadella
The way I come at it, Dwarkesh, it's a great question because at some level, if you're going to have this explosion, abundance, whatever, commodity of intelligence available, the first thing we have to observe is GDP growth.

Before I get to what Microsoft's revenue will look like, there's only one governor in all of this. This is where we get a little bit ahead of ourselves with all this AGI hype. Remember the developed world, which is what? 2% growth and if you adjust for inflation it's zero?

So in 2025, as we sit here, I'm not an economist, at least I look at it and say we have a real growth challenge. So, the first thing that we all have to do is, when we say this is like the Industrial Revolution, let's have that Industrial Revolution type of growth.

That means to me, 10%, 7%, developed world, inflation-adjusted, growing at 5%. That's the real marker. It can't just be supply-side.

In fact that’s the thing, a lot of people are writing about it, and I'm glad they are, which is the big winners here are not going to be tech companies. The winners are going to be the broader industry that uses this commodity that, by the way, is abundant. Suddenly productivity goes up and the economy is growing at a faster rate. When that happens, we'll be fine as an industry.

But that's to me the moment. Us self-claiming some AGI milestone, that's just nonsensical benchmark hacking to me. The real benchmark is: the world growing at 10%.

Dwarkesh Patel
Okay, so if the world grew at 10%, the world economy is $100 trillion or something, if the world grew at 10%, that's like an extra $10 trillion in value produced every single year. If that is the case, you as a hyperscaler... It seems like $80 billion is a lot of money. Shouldn't you be doing like $800 billion?

If you really think in a couple of years, we could be really growing the world economy at this rate, and the key bottleneck would be: do you have the compute necessary to deploy these AIs to do all this work?

Satya Nadella
That is correct. But by the way, the classic supply side is, "Hey, let me build it and they’ll come." That's an argument, and after all we've done that, we've taken enough risk to go do it.

But at some point, the supply and demand have to map. That's why I'm tracking both sides of it. You can go off the rails completely when you are hyping yourself with the supply-side, versus really understanding how to translate that into real value to customers.

That's why I look at my inference revenue. That's one of the reasons why even the disclosure on the inference revenue... It's interesting that not many people are talking about their real revenue, but to me, that is important as a governor for how you think about it.

You're not going to say they have to symmetrically meet at any given point in time, but you need to have existence proof that you are able to parlay yesterday's, let’s call it capital, into today's demand, so that then you can again invest, maybe exponentially even, knowing that you're not going to be completely rate mismatched.

Dwarkesh Patel
I wonder if there's a contradiction in these two different viewpoints, because one of the things you've done wonderfully is make these early bets. You invested in OpenAI in 2019, even before there was Copilot and any applications.

If you look at the Industrial Revolution, these 6%, 10% build-outs of railways and whatever things, many of those were not like, "We've got revenue from the tickets, and now we're going to…"

Satya Nadella
There was a lot of money lost.

Dwarkesh Patel
That's true. So, if you really think there's some potential here to 10x or 5x the growth rate of the world, and then you're like, "Well, what is the revenue from GPT-4?"

If you really think that's the possibility from the next level up, shouldn't you just, "Let's go crazy, let's do the hundreds of billions of dollars of compute?" I mean, there's some chance, right?

Satya Nadella
Here’s the interesting thing, right? That's why even that balanced approach to the fleet, at least, is very important to me. It's not about building compute. It's about building compute that can actually help me not only train the next big model but also serve the next big model. Until you do those two things, you're not going to be able to really be in a position to take advantage of even your investment.

So, that's kind of where it's not a race to just building a model, it's a race to creating a commodity that is getting used in the world to drive… You have to have a complete thought, not just one thing that you’re thinking about.

And by the way, one of the things is that there will be overbuild. To your point about what happened in the dotcom era, the memo has gone out that, hey, you know, you need more energy, and you need more compute. Thank God for it. So, everybody's going to race.

In fact, it's not just companies deploying, countries are going to deploy capital, and there will be clearly… I'm so excited to be a leaser, because, by the way; I build a lot, I lease a lot. I am thrilled that I'm going to be leasing a lot of capacity in '27, '28 because I look at the builds, and I'm saying, "This is fantastic." The only thing that's going to happen with all the compute builds is the prices are going to come down.

Dwarkesh Patel
Speaking of prices coming down, you recently tweeted after the DeepSeek model came out about Jevons' Paradox. I'm curious if you can flesh that out. Jevons' Paradox occurs when the demand for something is highly elastic. Is intelligence that bottlenecked on prices going down?

Because when I think about, at least my use cases as a consumer, intelligence is already so cheap. It's like two cents per million tokens. Do I really need it to go down to 0.02 cents? I'm just really bottlenecked on it becoming smarter. If you need to charge me 100x, do a 100x bigger training run. I'm happy for companies to take that.

But maybe you're seeing something different on the enterprise side or something. What is the key use case of intelligence that really requires it to get to 0.002 cents per million tokens?

Satya Nadella
I think the real thing is the utility of the tokens. Both need to happen: One is intelligence needs to get better and cheaper. And anytime there's a breakthrough, like even what DeepSeek did, with the efficient frontier of performance per token changes, the curve gets bent, and the frontier moves. That just brings more demand. That's what happened with cloud.

Here's an interesting thing: We used to think "oh my God, we've sold all the servers in the client-server era". Except once we started putting servers in the cloud, suddenly people started consuming more because they could buy it cheaper, and it was elastic, and they could buy it as a meter versus a license, and it completely expanded.

I remember going, let's say, to a country like India and talking about "here is SQL Server". We sold a little, but man, the cloud in India is so much bigger than anything that we were able to do in the server era. I think that's going to be true.

If you think about, if you want to really have, in the Global South, in a developing country, if you had these tokens that were available for healthcare that were really cheap, that would be the biggest change ever.

Dwarkesh Patel
I think it's quite reasonable for somebody to hear people like me in San Francisco and think "they're kind of silly; they don't know what it's actually like to deploy things in the real world".

As somebody who works with these Fortune 500s and is working with them to deploy things for hundreds of millions, billions of people, what's your sense on how fast deployment of these capabilities will be?

Even when you have working agents, even when you have things that can do remote work for you, with all the compliance and with all the inherent bottlenecks, is that going to be a big bottleneck, or is that going to move past pretty fast?

Satya Nadella
It is going to be a real challenge because the real issue is change management or process change. Here's an interesting thing: one of the analogies I use is, just imagine how a multinational corporation like us did forecasts pre-PC, and email, and spreadsheets. Faxes went around. Somebody then got those faxes and did an interoffice memo that then went around, and people entered numbers, and then ultimately a forecast came, maybe just in time for the next quarter.

Then somebody said, "Hey, I'm just going to take an Excel spreadsheet, put it in email, send it around. People will go edit it, and I'll have a forecast." So, the entire forecasting business process changed because the work artifact and the workflow changed.

That is what needs to happen with AI being introduced into knowledge work. In fact, when we think about all these agents, the fundamental thing is there's a new work and workflow.

For example, even prepping for our podcast, I go to my copilot and I say, "Hey, I'm going to talk to Dwarkesh about our quantum announcement and this new model that we built for game generation. Give me a summary of all the stuff that I should read up on before going." It knew the two Nature papers, it took that. I even said, "Hey, go give it to me in a podcast format." And so, it even did a nice job of two of us chatting about it.

So that became—and in fact, then I shared it with my team. I took it and put it into Pages, which is our artifact, and then shared. So the new workflow for me is I think with AI and work with my colleagues.

That's a fundamental change management of everyone who's doing knowledge work, suddenly figuring out these new patterns of "How am I going to get my knowledge work done in new ways?" That is going to take time. It's going to be something like in sales, and in finance, and supply chain.

For an incumbent, I think that this is going to be one of those things where—you know, let's take one of the analogies I like to use is what manufacturers did with Lean. I love that because, in some sense, if you look at it, Lean became a methodology of how one could take an end-to-end process in manufacturing and become more efficient. It's that continuous improvement, which is reduce waste and increase value.

That's what's going to come to knowledge. This is like Lean for knowledge work, in particular. And that's going to be the hard work of management teams and individuals who are doing knowledge work, and that's going to take its time.

Dwarkesh Patel
Can I ask you just briefly about that analogy? One of the things Lean did is physically transform what a factory floor looks like. It revealed bottlenecks that people didn't realize until you're really paying attention to the processes and workflows.

You mentioned briefly what your own workflow—how your own workflow has changed as a result of AIs. I'm curious if we can add more color to what will it be like to run a big company when you have these AI agents that are getting smarter and smarter over time?

Satya Nadella
It's interesting you ask that. I was thinking, for example, today if I look at it, we are very email heavy. I get in in the morning, and I'm like, man my inbox is full, and I'm responding, and so I can't wait for some of these Copilot agents to automatically populate my drafts so that I can start reviewing and sending.

But I already have in Copilot at least ten agents, which I query them different things for different tasks. I feel like there's a new inbox that's going to get created, which is my millions of agents that I'm working with will have to invoke some exceptions to me, notifications to me, ask for instructions.

So at least what I'm thinking is that there's a new scaffolding, which is the agent manager. It's not just a chat interface. I need a smarter thing than a chat interface to manage all the agents and their dialogue.

That's why I think of this Copilot, as the UI for AI, is a big, big deal. Each of us is going to have it. So basically, think of it as: there is knowledge work, and there's a knowledge worker. The knowledge work may be done by many, many agents, but you still have a knowledge worker who is dealing with all the knowledge workers. And that, I think, is the interface that one has to build.

Dwarkesh Patel
You're one of the few people in the world who can say that you have access to 200,000... you have this swarm of intelligence around you in the form of Microsoft the company and all its employees. And you have to manage that, and you have to interface with that, how to make best use of that. Hopefully, more of the world will get to have that experience in the future.

I'd be curious about how your inbox, if that means everybody's inbox, will look like yours in the morning.

Okay, before we get to that, I want to keep asking you more about AI, but I really want to ask you about the big breakthrough in quantum that Microsoft Research has announced. So can you explain what's going on?

Satya Nadella
This has been another 30-year journey for us. It's unbelievable. I'm the third CEO of Microsoft who's been excited about quantum.

The fundamental breakthrough here, or the vision that we've always had is, you need a physics breakthrough in order to build a utility-scale quantum computer that works. We took the path of saying, the one way for having a less noisy or more reliable qubit is to bet on a physical property that by definition is more reliable and that's what led us to the Majorana zero modes, which was theorized in the 1930s. The question was, can we actually physically fabricate these things? Can we actually build them?

So the big breakthrough effectively, and I know you talked to Chetan, was that we now finally have existence proof and a physics breakthrough of Majorana zero modes in a new phase of matter effectively. This is why we like the analogy of thinking of this as the transistor moment of quantum computing, where we effectively have a new phase, which is the topological phase, which means we can even now reliably hide the quantum information, measure it, and we can fabricate it. And so now that we have it, we feel like with that core foundational fabrication technique out of the way, we can start building a Majorana chip.

That Majorana One which I think is going to basically be the first chip that will be capable of a million qubits, physical. And then on that, thousands of logical qubits, error-corrected. And then it's game on. You suddenly have the ability to build a real utility-scale quantum computer, and that to me is now so much more feasible. Without something like this, you will still be able to achieve milestones, but you'll never be able to build a utility-scale computer. That's why we're excited about it.

Dwarkesh Patel
Amazing. And by the way, I believe this is it right here.

Satya Nadella
That is it.

Dwarkesh Patel
Yes.

Satya Nadella
I forget now, are we calling it Majorana? Yes, that's right. Majorana One. I'm glad we named it after that.

To think that we are able to build something like a million-qubit quantum computer in a thing of this size is just unbelievable. That's the crux of it: unless and until we could do that, you can't dream of building a utility-scale quantum computer.

Dwarkesh Patel
And you're saying the eventual million qubits will go on a chip this size? Okay, amazing.

Other companies have announced 100 physical qubits, Google's, IBM's, others. When you say you've announced one, but you're saying that yours is way more scalable in the limit.

Satya Nadella
Yeah. The one thing we’ve also done is we've taken an approach where we've separated our software and our hardware. We're building out our software stack, and we now have, with the neutral atom folks, the ion trap folks, we're also working with others who even have pretty good approaches with photonics and what have you, that means there'll be different types of quantum computers. In fact, we have what, I think that the last thing that we announced was 24 logical qubits. So we have also got some fantastic breakthroughs on error correction and that's what is allowing us, even on neutral atom and ion trap quantum computers, to build these 20 plus, and I think that'll keep going even throughout the year; you'll see us improve that yardstick.

But we also then said, "Let's go to the first principles and build our own quantum computer that is betting on the topological qubit." And that's what this breakthrough is about.

Dwarkesh Patel
Amazing. The million topological qubits, thousands of logical qubits, what is the estimated timeline to scale up to that level? What does the Moore's law here, if you've got the first transistor, look like?

Satya Nadella
We've obviously been working on this for 30 years. I'm glad we now have the physics breakthrough and the fabrication breakthrough.

I wish we had a quantum computer because by the way, the first thing the quantum computer will allow us to do is build quantum computers, because it's going to be so much easier to simulate atom-by-atom construction of these new quantum gates.

But in any case, the next real thing is, now that we have the fabrication technique, let us go build that first fault-tolerant quantum computer. And that will be the logical thing.

So, I would say now I can say, "Oh, maybe '27, '28, '29, we will be able to actually build this." Now that we have this one gate, can I put the thing into an integrated circuit and then actually put these integrated circuits into a real computer? That is where the next logical step is.

Dwarkesh Patel
And what do you see as, in '27, '28, you've got it working? Is it a thing you access through the API? Is it something you're using internally for your own research in materials and chemistry?

Satya Nadella
It’s a great question. One thing that I've been excited about is, even in today's world... we had this quantum program, and we added some APIs to it. The breakthrough we had maybe two years ago was to think of this HPC stack, and AI stack, and quantum together.

In fact, if you think about it, AI is like an emulator of the simulator. Quantum is like a simulator of nature. What is quantum going to do? By the way, quantum is not going to replace classical. Quantum is great at what quantum can do, and classical will also...

Quantum is going to be fantastic for anything that is not data-heavy but is exploration-heavy in terms of the state space. It should be data-light but exponential states that you want to explore. Simulation is a great one: chemical physics, what have you, biology.

One of the things that we've started doing is really using AI as the emulation engine. But you can then train. So the way I think of it is, if you have AI plus quantum, maybe you'll use quantum to generate synthetic data that then gets used by AI to train better models that know how to model something like chemistry or physics or what have you. These two things will get used together.

So even today, that's effectively what we're doing with the combination of HPC and AI. I hope to replace some of the HPC pieces with quantum computers.

Dwarkesh Patel
Can you tell me a little bit about how you make these research decisions which, in 20 years time, 30 years time, will actually pay dividends, especially at a company of Microsoft's scale? Obviously, you're in great touch with the technical details in this project. Is it feasible for you to do that with all the things Microsoft Research does?

How do you know the current bet you're making will pay out in 20 years? Does it just have to emerge organically through the org, or how are you keeping track of all this?

Satya Nadella
The thing that I feel was fantastic is when Bill, when he started MSR back in '95 I guess. I think in the long history of these curiosity-driven research organizations, to just do a research org that is about fundamental research and MSR, over the years, has built up that institutional strength so when I think about capital allocation or budgets, we first put the chips in and say, "Here is MSR's budget." We gotta go at it each year knowing that most of these bets are not going to pay off in any finite time frame. Maybe the sixth CEO of Microsoft will benefit from it. And in tech that is I think a given.

The real thing that I think about is, when the time has come for something like quantum or a new model or what have you, can you capitalize? So as an incumbent, if you look at the history of tech, it's not that people didn't invest. It's that you need to have a culture that knows how to take an innovation and scale it.

That's the hard part, quite frankly, for CEOs and management teams. Which is kind of fascinating. It's as much about good judgment as it is about good culture. Sometimes we've gotten it right; sometimes we've gotten it wrong; I can tell you the thousand projects from MSR that we should have probably led with, but we didn't. And I always ask myself why. It's because we were not able to get enough conviction and that complete thought of how to not only take the innovation but make it into a useful product with a business model that we can then go to market with.

That's the job of CEOs and management teams: not to just be excited about any one thing, but to be able to actually execute on a complete thing. And that's easier said than done.

Dwarkesh Patel
When you mentioned the possibility of three subsequent CEOs of Microsoft, if each of them increases the market cap by an order of magnitude, by the time you've got the next breakthrough, you'll be like the world economy or something.

Satya Nadella
Or remember, the world is going to be growing at 10%, so we'll be fine.

Dwarkesh Patel
Let's dig into the other big breakthrough you've just made. It's amazing that you have both of them coming out the same day, in your gaming world models. I'd love if you can tell me a little bit about that.

Satya Nadella
We're going to call it Muse. It's going to be the model of this world action, or human action model.

This is very cool. One of the things is that obviously, Dall-E and Sora have been unbelievable in what they've been able to do in terms of generative models. One thing that we wanted to go after was using gameplay data. Can you actually generate games that are both consistent and then have the ability to generate the diversity of what that game represents, and then are persistent to user mods?

That's what this is. They were able to work with one of our game studios, and this is the other publication in Nature.

The cool thing is what I'm excited about is bringing—we're going to have a catalog of games soon that we will start using these models, or we're going to train these models to generate, and then start playing them.

In fact, when Phil Spencer first showed it to me, he had an Xbox controller and this model basically took the input and generated the output based on the input. And it was consistent with the game. That to me is a massive moment of “wow”. It's kind of like the first time we saw ChatGPT complete sentences, or Dall-E draw, or Sora. This is one such moment.

Dwarkesh Patel
I got a chance to see some of the videos in the real-time demo this morning with your lead researcher Katja on this. Only once I talked to her did it really hit me how incredible this is, in the sense that we've used AI in the past to model agents, and just using that same technique to model the world around the agent gives consistent real-time – we'll superimpose videos of what this looks like atop this podcast so people can get a chance to see it for themselves. I guess it'll be out by then, so they can also watch it there.

This in itself is incredible. You, through your span as CEO, have invested tens of hundreds of billions of dollars in building up Microsoft Gaming and acquiring IP.

In retrospect, if you can just merge all of this data into one big model that can give you this experience of visiting and going through multiple worlds at the same time, and if this is the direction gaming is headed, it seems like a pretty good investment to have made. Did you have any premonition about this?

Satya Nadella
I wouldn't say that we invested in gaming to build models. We invested, quite frankly, because- here's an interesting thing about our history: We built our first game before we built Windows. Flight Simulator was a Microsoft product long before we even built Windows.

So, gaming has got a long history at the company, and we want to be in gaming for gaming's sake. I always start by saying I hate to be in businesses where they're means to some other end. They have to be ends unto themselves.

And then, yes, we're not a conglomerate. We are a company where we have to bring all these assets together and be better owners by adding value. For example, cloud gaming is a natural thing for us to invest in because that will just expand the TAM and expand the ability for people to play games everywhere.

The same thing with AI and gaming: we definitely think that it can be helpful in maybe changing- it's kind of like the CGI moment, even for gaming long-term. And it's great. As the biggest, world's largest publisher, this will be helpful. But at the same time, we've got to produce great quality games. I mean, you can't be a gaming publisher without, sort of, first and foremost being focused on that.

But the fact that this data asset is going to be interesting, not just in a gaming context, but it's going to be a general action model and a world model, it's fantastic. I mean like, you know, I think about gaming data as perhaps, you know, what YouTube is perhaps to Google, gaming data is to Microsoft. And so therefore I'm excited about that.

Dwarkesh Patel
Yeah, and that's what I meant, just in the sense of like, you can have one unified experience across many different kinds of games. How does this fit into the other, separate from AI, the other things that Microsoft has worked on in the past, like mixed reality? Maybe giving smaller game studios a chance to build these AAA action games? Just like five, ten years from now, what kinds of ways could you imagine?

Satya Nadella
I've thought about these three things as the cornerstones of, in an interesting way, even five, six, seven years ago is when I said the three big bets that we want to place [are] AI, quantum, and mixed reality. And I still believe in them, because in some sense, what are the big problems to be solved?

Presence. That's the dream of mixed reality. Can you create real presence? Like you and I doing a podcast like this.

I think it’s still proving to be the harder one of those challenges, quite honestly. I thought it was going to be more solvable. It's tougher, perhaps, just because of the social side of it: wearing things and so on.

We're excited about, in fact, what we're going to do with Anduril and Palmer, now, with even how they'll take forward the IVAS program, because that's a fantastic use case. And so we'll continue on that front.

But also, the 2D surfaces. It turns out things like Teams, right, thanks to the pandemic, we've really gotten the ability to create essentially presence through even 2D. And that I think will continue. That's one secular piece.

Quantum we talked about, and AI is the other one. So these are the three things that I look at and say, how do you bring these things together? Ultimately, not as tech for tech's sake, but solving some of the fundamental things that we, as humans, want in our life, and more, we want them in our economy, driving our productivity. And so if we can somehow get that right, then I think we will have really made progress.

Dwarkesh Patel
When you write your next book, you've got to have some explanation of why those three pieces all came together around the same time, right? Like, there's no intrinsic reason you would think quantum and AI should happen in 2028 and 2025 and so forth.

Satya Nadella
That's right. At some level, I look at it and say: the simple model I have is, hey is there a systems breakthrough? To me, the systems breakthrough is the quantum thing.

Is there a business logic breakthrough? That's AI to me, which is: can the logic tier be fundamentally reasoned differently? Instead of imperatively writing code, can you have a learning system? That's the AI one.

And then the UI side of it is presence.

Dwarkesh Patel
Going back to AI for a second, in your 2017 book... 2019 you invest in OpenAI, very early, 2017 is even earlier, you say in your book, "One might also say that we're birthing a new species, one whose intelligence may have no upper limits."

Now, super-early, of course, to be talking about this in 2017. We've been talking in a granular fashion about agents, Office Copilot, capex, and so forth. But if you zoom out and consider this statement you've made, and you think about you as a hyperscaler, as the person doing research in these models as well, providing training, inference, and research for building a new species, how do you think about this in the grand scheme of things?

Do you think we're headed towards superhuman intelligence in your time as CEO?

Satya Nadella
I think even Mustafa uses that term. In fact he’s used that term more recently, this “new species”.

The way I come at it is, you definitely need trust. Before we claim it is something as big as a species, the fundamental thing that we've got to get right is that there is real trust, whether it's personal or societal level trust, that's baked in. That's the hard problem.

I think the one biggest rate limiter to the power here will be how does our legal… call it infrastructure, we’re talking about all the compute infrastructure, well how does the legal infrastructure evolve to deal with this? This entire world is constructed with things like humans owning property, having rights, and being liable. That’s the fundamental thing that one has to first say, okay what does that mean for anything that now humans are using as tools? And if humans are going to delegate more authority to these things, then how does that structure evolve? Until that really gets resolved, I don't think just talking about the tech capability is going to happen.

Dwarkesh Patel
As in, we won't be able to deploy these kinds of intelligences until we figure out how to…?

Satya Nadella
Absolutely. Because at the end of the day, there is no way. Today, you cannot deploy these intelligences unless and until there's someone indemnifying it as a human.

To your point, I think that's one of the reasons why I think about even the most powerful AI is essentially working with some delegated authority from some human. You can say, oh, that's all alignment and this, that, and the other. That's why I think you have to really get these alignments to work and be verifiable in some way, but I just don't think that you can deploy intelligences that are out of control. For example, this AI takeoff problem may be a real problem, but before it is a real problem, the real problem will be in the courts. No society is going to allow for some human to say, "AI did that."

Dwarkesh Patel
Yes. Well, there's a lot of societies in the world, and I wonder if any one of them might not have a legal system that might be more amenable. And if you can't have a takeoff, then you might worry. It doesn't have to happen in America, right?

Satya Nadella
We think that no society cares about it, right? There can be rogue actors, I'm not saying there won't be rogue actors; there are cyber criminals and rogue states; they're going to be there.

But to think that human society at large doesn't care about it is also not going to be true. I think we all will care. We know how to deal with rogue states and rogue actors today. The world doesn't sit around and say “we’ll tolerate that”. That's why I'm glad that we have a world order in which anyone who is a rogue actor in a rogue state has consequences.

Dwarkesh Patel
Right. But if you have this picture where you can have 10% economic growth, I think it really depends on getting something like AGI working, because tens of trillions of dollars of value, that sounds closer to the total of human wages, around $60 trillion of the economy. Getting that magnitude, you kind of have to automate labor or supplement labor in a very significant way.

If that is possible, and once we figure out the legal ramifications for it, it seems quite plausible, even within your tenure that we figure that out. Are you thinking about superhuman intelligence? Like, the biggest thing you do in your career is this?

Satya Nadella
You bring up another point. I know David Autor and others have talked a lot about this which is, 60% of labor- I think the other question that needs to happen, let’s at least talk about our democratic societies. I think that in order to have a stable social structure, and democracies function, you can’t just have a return on capital and no return on labor. We can talk about it, but that 60% has to be revalued.

In my own simple way, maybe you can call it naive, we'll start valuing different types of human labor. What is today considered high-value human labor may be a commodity. There may be new things that we will value.

Including that person who comes to me and helps me with my physical therapy or whatever, whatever is going to be the case that we value, but ultimately, if we don't have return on labor, and there's meaning in work and dignity in work and all of that, that's another rate limiter to any of these things being deployed.

Dwarkesh Patel
On the alignment side, two years ago, you guys released Sydney Bing. Just to be clear, I think given the level of capabilities at the time, it was a charming, endearing, kind of funny example of misalignment.

But that was because, at the time, it was like chatbots. They can go think for 30 seconds and give you some funny or inappropriate response. But if you think about that kind of system—that, I think to a New York Times reporter, tried to get him to leave his wife or something—if you think about that going forward, and you have these agents that are for hours, weeks, months going forward, just like autonomous swarms of AGIs, who could be in similar ways misaligned and screwing stuff up, maybe coordinating with each other, what's your plan going forward so that when you get the big one, you get it right?

Satya Nadella
That is correct. That's one of the reasons why when we usually allocate compute, let's allocate compute for what is that alignment challenge?

And then more importantly, what is the runtime environment in which you are really going to be able to monitor these things? The observability around it? We do deal with a lot of these things today in the classical side of things as well, like cyber. We don't just write software and then just let it go. You have software and then you monitor it. You monitor it for cyber attacks, you monitor it for fault injections, and what have you.

Therefore, I think we will have to build enough software engineering around the deployment side of these, and then inside the model itself, what's the alignment? These are all, some of them are real science problems. Some of them are real engineering problems, and then we will have to tackle it.

That also means taking our own liability in all of this. So that's why I'm more interested in deploying these things in where you can actually govern what the scope of these things is, and the scale of these things is. You just can't unleash something out there in the world that creates harm, because the social permission for that is not going to be there.

Dwarkesh Patel
When you get the agents that can really just do weeks worth of tasks for you, what is the minimum assurance you want before you can let it run a random Fortune 500?

Satya Nadella
I think when I use something like Deep Research, even, the minimum assurance I think we want is before we especially have physical embodiment of anything, that I think is kind of one of those thresholds, when you cross. That might be one place.

Then the other one is, for example, the permissions of the runtime environment in which this is operating. You may want guarantees that it's sandboxed, it is not going out of that sandbox.

Dwarkesh Patel
I mean, we already have web search and we already have it out of the sandbox.

Satya Nadella
But even what it does with web search and what it writes — for example to your point, if it's just going to write a bunch of code in order to do some computation, where is that code

deployed? And is that code ephemeral for just creating that output, versus just going and springing that code out into the world?

Those are things that you could, in the action space, actually go control.

Dwarkesh Patel
And separate from the safety issues, as you think about your own product suite, and you think about, if you do have AIs this powerful, at some point, it's not just like Copilot- an example you mentioned about how you were prepping for this podcast- it's more similar to how you actually delegate work to your colleagues.

What does it look like, given your current suite, to add that in? I mean, there's one question about whether LLMs get commodified by other things.

I wonder if these databases or canvases or Excel sheets or whatever — if the LLM is your main gate point into accessing all these things, is it possible that the LLMs commodify Office?

Satya Nadella
It's an interesting one. The way I think about the first phase, at least, would be: Can the LLM help me do my knowledge work using all of these tools or canvases more effectively?

One of the best demos that I've seen is a doctor getting ready for a tumor board workflow. She's going into a tumor board meeting, and the first thing she uses Copilot for is to create an agenda for the meeting because the LLM helps reason about all the cases, which are in some SharePoint site. It says, "Hey, these cases — obviously, a tumor board meeting is a high-stakes meeting where you want to be mindful of the differences in cases so that you can then allocate the right time."

Even that reasoning task of creating an agenda that knows how to split time- super. So, I use the LLM to do that. Then I go into the meeting, I'm in a Teams call with all my colleagues. I'm focused on the actual case versus taking notes, because you now have this AI copilot doing a full transcription of all of this. It's not just a transcript, but a database entry of what is in the meeting that is recallable for all time.

Then she comes out of the meeting, having discussed the case and not been distracted by note-taking. She's a teaching doctor; she wants to go and prep for her class. And so she goes into Copilot and says, "Take my tumor board meeting and create a PowerPoint slide deck out of it so that I can talk to my students about it."

So that’s the type. The UI and the scaffolding that I have are canvases that are now getting populated using LLMs. And the workflow itself is being reshaped; knowledge work is getting done.

Here's an interesting thing: If someone came to me in the late '80s and said, "You're going to have a million documents on your desk," I would say, "What the heck is that?" I would have literally thought there was going to be a million physical copies of things on my desk. Except, we do have a million spreadsheets and a million documents.

Dwarkesh Patel
I don’t, you do.

Satya Nadella
They're all there. And so, that's what's going to happen with even agents. There will be a UI layer. To me, Office is not just about the office of today; it's the UI layer for knowledge work. It'll evolve as the workflows evolve. That's what we want to build.

I do think the SaaS applications that exist today, these CRUD applications, are going to fundamentally be changed because the business logic will go more into this agentic tier. In fact, one of the other cool things today in my Copilot experience is when I say, "Hey, I'm getting ready for a meeting with a customer," I just go and say, "Give me all the notes for it that I should know." It pulls from my CRM database, it pulls from my Microsoft Graph, creates a composite, essentially artifact, and then it applies even logic on it. That, to me, is going to transform the SaaS applications as we know of it today.

Dwarkesh Patel
SaaS as an industry might be worth hundreds of billions to trillions of dollars a year, depending on how you count. If really that can just get collapsed by AI, is the next step up in your next decade 10X-ing the market cap of Microsoft again? Because you're talking about trillions of dollars…

Satya Nadella
It would also create a lot of value in the SaaS. One thing we don't pay as much attention to perhaps is the amount of IT backlog there is in the world.

These code gen things, plus the fact that I can interrogate all of your SaaS applications using agents and get more utility will be the greatest explosion of apps, they'll be called agents, so that for every vertical, in every industry, in every category, we're suddenly going to have the ability to be serviced.

So there's going to be a lot of value. You can't stay still. You can't just say the old thing of, "Oh, I schematized some narrow business process, and I have a UI in the browser, and that's my thing." That's ain’t going to be the case. You have to go up-stack and say, "What's the task that I have to participate in?"

You will want to be able to take your SaaS application and make it a fantastic agent that participates in a multi-agent world. As long as you can do that, then I think you can even increase the value.

Dwarkesh Patel
Can I ask you some questions about your time at Microsoft?

Satya Nadella
Yeah.

Dwarkesh Patel
Is being a company man underrated? So you've spent most of your career at Microsoft, and you could say that one of the reasons you've been able to add so much value is you've seen the culture, the history, and the technology. You have all this context by rising up through the ranks. Should more companies be run by people who have this level of context?

Satya Nadella
That's a great question. I've not thought about it that way.

Through my 34 years now of Microsoft, each year I felt more excited about being at Microsoft versus thinking that, oh, I'm a company person or what have you. I take that seriously, even for anybody joining Microsoft. It's not like they're joining Microsoft as long as they feel that they can use this as a platform for their both economic return, but also a sense of purpose and a sense of mission that they can accomplish by using us as a platform. That's the contract.

So I think yes, companies have to create a culture that allows people to come in and become company people like me. Microsoft got it more right than wrong, at least in my case, and I hope that remains the case.

Dwarkesh Patel
The sixth CEO that you’re talking about, who’ll get to use the research you’re starting now, what are you doing to retain the future Satya Nadellas so that they're in a position to become future leaders?

Satya Nadella
It's fascinating. This is our 50th year, and I think a lot about it. The way to think about it is, longevity is not a goal; relevance is.

The thing that I have to do and all 200,000 of us have to do every day is: Are we doing things that are useful and relevant for the world as we see it evolving, not just today, but tomorrow?

We live in an industry where there's no franchise value, so that’s the other hard part. If you take the R&D budget that we will spend this year, it’s all speculation on what's going to happen five years from now. You have to basically go in with that attitude, saying, "We are doing things that we think are going to be relevant."

So that's what you have to focus on. Then know that there's a batting average, and you're not going to get- you have to have a high tolerance for failure. You have to take enough shots on goal to be able to say, "Okay, we will make it to the other side as a company." That's what makes it tricky in this industry.

Dwarkesh Patel
Speaking of - you just mentioned that you're two months away from your 50th anniversary of Microsoft’s founding. If you look at the top 10 companies by market cap, or top 5, basically, everybody else but Microsoft is younger than Microsoft. It's an interesting observation about why the most successful companies often are quite young. The average Fortune 500 company will last 10 to 15 years.

What has Microsoft done to remain relevant for this many years? How do you keep refounding?

Satya Nadella
I love that, Reed Hoffman uses that term, "refounding." That's the mindset. People talk about founder mode, but for us mere mortal CEOs, it's more like refounder mode.

To be able to see things again in a fresh way is the key. To your question: can we culturally create an environment where refounding becomes a habit thing? Every day we come in and say, "We feel we have a stake in this place to be able to change the core assumptions of what we do and how we relate to the world around us. Do we give ourselves permission?” I think many times, companies feel over-constrained by either business model or whatever. You just have to unconstrain yourself.

Dwarkesh Patel
If you did leave Microsoft, what company would you start?

Satya Nadella
Company I would start? Man. That’s where the company man and me sort of says, “I'll never leave Microsoft.”

If I were thinking of doing something, I think picking a domain that has… When I look at the dream of tech, we've always said technology is about the biggest, greatest democratizing force.

I feel like finally, we have that ability. If you say those tokens per dollar per watt is what we can generate, I would love to find some domain in which that can be applied, where it is so underserved.

That's where healthcare, education… Public sector would be another place. If you take those domains, which are the underserved places, where my life as a citizen of this country or a member of this society or anywhere, would I be better off if somehow all this abundance translated into better healthcare, better education, and better public sector institutions serving me as a citizen? That would be a place.

Dwarkesh Patel
One thing I'm not sure about, hearing your answers on different questions, is whether you think AGI is a thing. Will there be a thing which automates all cognitive labor, like anything anybody can do on a computer?

Satya Nadella
This is where I have a problem with the definitions of how people talk about it. Cognitive labor is not a static thing. There is cognitive labor today. If I have an inbox that is managing all my agents, is that new cognitive labor?

Today's cognitive labor may be automated. What about the new cognitive labor that gets created? Both of those things have to be thought of, which is the shifting…

That's why I make this distinction, at least in my head: Don't conflate knowledge worker with knowledge work. The knowledge work of today could probably be automated. Who said my life's goal is to triage my email? Let an AI agent triage my email.

But after having triaged my email, give me a higher-level cognitive labor task of, "Hey, these are the three drafts I really want you to review." That's a different abstraction.

Dwarkesh Patel
But will AI ever get to the second thing?

Satya Nadella
It may, but as soon as it gets to that second thing, there will be a third thing. Why are we thinking that somehow, when we have dealt with tools that have changed what cognitive labor is in history, why are we worried that all cognitive labor will go away?

Dwarkesh Patel
I'm sure you've heard these examples before, but the idea that horses can still be good for certain things, there are certain terrains you can't take a car on. But the idea that you're going to see horses around the street, they're going to employ millions of horses, it's just not happening.

And then the idea is, could a similar thing happen with humans?

Satya Nadella
But in one very narrow dimension? It's only 200 years of history of humans where we have valued some narrow sort of things called "cognitive labor" as we understand it.

Let's take something like chemistry. If this thing, quantum plus AI really helped us do a lot of novel material science and so on, that's fantastic to have novel material science being done by it. Does that take away from all the other things that humans can do?

Why can't we exist in a world where there are powerful cognitive machines, knowing that our cognitive agency has not been taken away?

Dwarkesh Patel
I'll ask this question, not about you, but in a different scenario, so maybe you can answer it without embarrassment. Suppose on the Microsoft board, could you ever see adding an AI to the board? Could it ever have the judgment, context, and holistic understanding to be a useful advisor?

Satya Nadella
It's a great example. One of the things we added was a facilitator agent in Teams. The goal there, it's in the early stages, is can that facilitator agent use long-term memory, not just on the context of the meeting, but with the context of projects I'm working on, and the team, and what have you, be a great facilitator?

I would love it even in a board meeting, where it's easy to get distracted. After all, board members come once a quarter, and they're trying to digest what is happening with a complex company like Microsoft. A facilitator agent that actually helped human beings all stay on topic and focus on the issues that matter, that's fantastic.

That's kind of literally having, to your point about even going back to your previous question, having something that has infinite memory that can even help us. You know, after all, what is that Herbert Simon thing? We are all bounded rationality. So if the bounded rationality of humans can actually be dealt with because there is a cognitive amplifier outside, that's great.

Dwarkesh Patel
Speaking of materials and chemistry stuff, I think you said recently that you want the next 250 years of progress in those fields to happen in the next 25 years. Now, when I think about what's going to be possible in the next 250 years, I'm thinking like space travel, and space elevators, and immortality, and curing all diseases. Next 25 years, you think?

Satya Nadella
One of the reasons why I brought that up was, I love that thing of, the industrial revolution was the 250 years. We have to take this entire change from a carbon-based system to something different.

That means you have to fundamentally reinvent all of what has happened with chemistry over the last 250 years. That's where I hope we have this quantum computer, this quantum computer helps us get to new materials, and then we can fabricate those new materials that help us with all of the challenges we have on this planet. And then I'm all for interplanetary travel.

Dwarkesh Patel
Amazing. Satya, thank you so much for your time.

Satya Nadella
Thank you so much. It's wonderful. Thanks.

Dwarkesh Patel
Great, thank you.