Konstantine Buhler: Good morning, everyone. My name is Konstantine Buhler, and I'm a partner at Sequoia Capital, focused on AI investing. Both Nvidia and Citadel Securities actually have a lot in common. They're both exceptional businesses. Really well-run. Oh, really brilliant leaders. They are exceptionally well-run. They were both powered by the computing revolution, and both are leaders in their respective industries with technology. They also have another, lesser-known fact. In both cases, their first outside investor was Sequoia Capital. So.

Jensen Huang: $1 million. They, they risked $1 million in Nvidia in 1993.

Buhler: You were worth it.

Huang: One solid million dollars. Went way out on the limb.

Buhler It was a little more into Citadel Securities. So, when we were asked to speak about AI at this conference, it was incredibly clear who the best person in the world to speak would be. It's the man who has built the entire infrastructure for the AI revolution upon which all of this AI rests, and the man who built the most valuable company in the world. Please join me in welcoming Jensen Huang.

Huang: Oh, thanks. It's nice to wake up this way.

Buhler: But you've been working for hours.

Huang: Yes, I have.

Buhler: So, Jensen, we have a room full of institutional investors who are some of the best in the world. They manage many trillions of AUM, and they are constantly looking for edge. You are someone who always has edge. And every one of our conversations, you have compelling insights about what the future is going to look like. In the next 60 minutes, we have an ambitious agenda to cover stories of the edge from the very beginning of Nvidia all the way through its rise through the center of the AI revolution. And then we'll spend the majority of the time on what's next for Nvidia and AI. OK. So let's start at the very beginning. It's 1993. You're 30 years old. What's the insight that gave you the edge to start Nvidia?

Huang: We were going through, basically, the PC revolution and the revolution of CPUs. It was the era of Moore's Law. It was the the time when, integrated microprocessors, Intel, Moore's law, the scaling, the scaling laws of transistors. That was the buzz. And that was nearly all of the investment dollars in Silicon Valley. And the computer industry. And, and, we observed, something a little different. We said, "There are many problems that- one of the benefits of the CPUs is the general purpose. But the the fundamental problem of general purpose technologies is that they tend not to

be very good, extremely good at very hard problems." And so we conjecture two things. One, we observe that, that, there are problems that we could solve with an accelerator, that is more domain-specific, more targeted. And those problems could be interesting to solve. And we observed that the general-purpose technologies, the shrinking of these transistors would eventually reach the limits. The idea that you could keep reducing the size of transistors and scaling it using this, this technique, it is a, it's a set of heuristics called Dennard scaling and Dennard scaling. And, Mead and Conway came up with what are really the fundamental principles behind Moore's Law. And if you go back to those, you'll discover that there will be a limit to how far you can shrink transistors. And at someday, you'll get you'll get diminishing returns. And there are large computing problems. We believe that the computing problems that we could solve are nearly infinite in scale. And so, one of these days, a new type of computing approach would emerge. And we, you know, we, focused our company on, on augmenting, supplementing, general-purpose computing with this technology called accelerated computing. I said, that was really the observation. And and so you said something earlier about how Nvidia is always ahead of the curve. Oftentimes, oftentimes, if you reason about things from first principles, what's, what's working today incredibly. Well, if you could reason about it from first principles and ask yourself, on what foundation is that first principle built on top, and how would that change over time? It allows you to, you know, hopefully, see around corners.

Buhler: So when you built the graphics accelerator, you're early to the party, but then hundreds of other competitors sprung up. Yeah. You eventually won in that market in the early 2000s. You said, "Hey, this technology might be able to generalize itself." You're talking about the generalization of a CPU. Perhaps the GPU could also be generalized for more processing. Yeah. Let's talk about CUDA. Yeah. How did that come about? Where did you get that insight? A story goes that it's from researchers. How did you read their work and conclude that the GPU could be a general computer?

Huang: Well, the, so first of all, the reason why Nvidia was hard to build was because we had to invent a new technology and invent a market. And at the time, in 1993, in order to create a new computing platform, you need a large market and Silicon Graphics, which were doing 3D graphics at the time, the markets are too small to enable a new computing platform. And, so if we wanted to create a new computing architecture, we need a new and large market. And that large market didn't exist because the architecture didn't exist. You got the chicken and the egg problem. And so what Nvidia became good at and the modern 3D graphics video game market, we contributed tremendously to. And so Sequoia Capital's big issue at the time with Nvidia's, funding principles that we had to go invent the technology and the market simultaneously. And the odds of that happening is approximately 0%. And I still

remember, when I pitched the story and I said, and, Don Valentine at the time. "Well, Jensen, where's your where's your app? Where's the killer app?" And I said, "Oh yeah, there's this company called Electronic Arts." And I didn't realize, Don had just invested in Electronic Arts. Electronic Arts, they're going to create, we're going to help them create 3D graphics games and where to create this market. And he goes, "You know, Jensen, I want you to know that we invested in Electronic Arts and their CTO is 14 years old and is driven to work. And you're telling me that's your killer app?" And so, anyhow, you know, we, we created the modern 3D graphics gaming ecosystem. And as you know, it's one of the largest entertainment industries in the world. The fundamental problem of 3D graphics is basically simulating reality. If you go back to first principles, what it's doing is trying to recreate reality and the, the, the, the fundamental, the mathematics of, of, reproducing photorealistic images and dynamic worlds is, you know, fundamentally physics simulation. And so, linear algebra is obviously very important to it. And, and, we realize that concept. And so the question is: How do you bring something general purpose into something very, very specialized? And that's the great, that's the great invention of our company. We invented the technology. we invented the market. And we invent invented the pathways for us to systematically grow from a very vertically focused industry to eventually become more and more general purpose. And so that hardly ever happens. And that pathway was hard to do. But I don't want to take up the rest of the time explaining it. But I think the CUDAS invention is part invention of the technology, which is observation, of how we can generalize our GPUs. But it's, a lot of it is about the invention of new products, how to take it to market, invention of new strategies, how to get the market to adopt it, and, eventually, an invention inventing essentially ecosystems that ultimately creates the flywheel that makes a computing platform happen. So we invented all of those things. They're all brand new. And, and if you go back, you just take a step back, and you ask yourself, aside from ARM and aside from x86, what is another computing platform that exists in the world that almost everybody uses? Doesn't exist. And so inventing a new computing platform rarely happens. And, in our case, it took us almost 30 years.

Buhler: So you were able to take this very specialized, extremely high-performant acceleration device and generalize it so that researchers and academics around the world would be able to run their processing much faster. You know, that the Moore's Law limitations that they were up against all of a sudden were relaxed dramatically. Yeah. Now, let's jump forward to the early 2010s. You know, at the time, deep learning was kind of an academic backwater. The idea of neural networks had gone through a winter phase. Then in 2012, there was a breakthrough with AlexNet in computer vision. And that was all accelerated on Nvidia GPUs. Was that the moment that you realized

this AI revolution was becoming real? And if so, how did you capitalize on it? What was the edge to make Nvidia the center of this revolution?

Huang: Yeah, two, two serendipitous, Two serendipitous moments. And then one, which is just a great, again, first principle observation about deep learning. The serendipity started with, I was trying to solve computer vision and, and, we wanted to solve computer vision for a lot of different reasons, you know? Anyways, we wanted to solve computer vision. And computer vision was really brittle, really hard to generalize, a collection of a whole bunch of tricks. And I really, really hated how the industry was evolving and really quite frustrated with the progress. Meanwhile, one of our major strategies of democratizing the architecture is to go to higher, to get scientists in higher education to use our platform. Use CUDA. And so I started with seismic processing, molecular dynamics, particle physics, you know, quantum chemistry. I went, you know, I took I took a Nvidia for CUDA everywhere. And actually, there was a, there was a strategy at the company called "CUDA everywhere" that that meant Jensen schlepping CUDA all over the world. And so I went to universities everywhere, and we meet with researchers and, and that initiative of getting CUDA into higher education and researchers everywhere caused some researchers to reach out to us, in 2012, 2011. And, Jeff Hinton was trying to solve computer vision. And Andrew Ng was trying to solve computer vision. And, Yann LeCun was trying to solve computer vision because there was this, there was a contest coming up called ImageNet that Fei-Fei is in charge of. And I was trying to solve computer vision. And so when you're naturally trying to solve a problem, and then all these interesting people are solving similar problems that attract your attention. So that's serendipity. The thing that that was great observation is that, we could create a new type of solver for them that's called Q DNN. Kind of like the sequel of in-storage computing. We invented Q DNN, which is, in-network computing, if you will, and that way of doing computation, there's a library called Q DNN, made it possible for all of them to use CUDA successfully. But the thing that was was I saw the same results as everybody else. You know, everybody saw the big jump in computer vision effectiveness. But where we took it further was we reasoned about, so, so, this is so good at computer vision and why and what else could it be good at? And the ability for, for, deep neural networks to be extremely deep, meaning because each layer is trained independently of the others. And you could backpropagate, from a loss function all the way back to its input. You could learn almost any function. And we came to the conclusion this is a universal function approximator. And if we can then add to it state, which is, you know, CNN was, kind of a two-dimensional, multi-dimensional, pattern recognizer. And then RNNs gave you a state machine within it, and LSTM gives you an even better state machine, and then transformers give you the ultimate state machine. And so, so the, the idea that that we would have a universal function approximator that can learn almost any function. Well

the question is, is what problem can't it solve? Now invert the question. And we came to the conclusion most of the problems we wanted to solve could have a deep learning component to it. And so we decided, you know, how would we reason about where deep learning could be 10 years from now? 20 years from now? We broke down the computation problem, and we came to the conclusion that every single chip, every single system, every software, every single layer of the computing stack could be reinvented. And that that decision to go after it was probably, you know, one of the better decisions in history.

Buhler: I was doing AI research at the time at Stanford, and the major constraint was always the compute. You know, we had limited clusters in order to run these algorithms. And Nvidia came in and not only relaxed that compute, but made it possible with with the CUDA infrastructure. You know, that is largely your history. You make more and more compute possible. In 2016, you very famously created the world's first AI factory, the DGX-1, You know, you actually hand delivered it to Elon Musk at OpenAI. I built this brand new computer, and it's, it doesn't, it doesn't look like anything the world's ever seen before. It doesn't work like anything the world's ever seen before. And I remember announcing it at GTC, and literally the audience was just like this. Nobody knew what I was talking about. And. That was a joke. And so, so with the same amount of applause. and so, so I announce this thing and Everybody goes "Uh huh." And literally on that GTC, I was on, I invited Elon to talk about something. The two of us were working on self-driving cars. And so he came on stage. He says "Jensen, what's that computer?" "That's DGX-1, and I built it for this reason." And he goes, "I could use one." And I finally got a P0. And then he goes, he goes, he goes, "Yeah, I have this nonprofit." That's ... [labored sigh] Oh, you know, when you build something brand new, the last thing you want to hear is your first customer is a nonprofit. And so anyhow, anyhow, I delivered, I delivered, I was the DoorDash computer guy, and I DoorDashed this computer up to San Francisco, and the company was OpenAI.

Buhler: It is a very profitable nonprofit or revenue-scale nonprofit.

Huang: We've been working together for a long time. Every, every, every model has been built on Nvidia since. Yeah.

Buhler: And this thing is huge physically. When Jensen talking about a computer, we're talking about a massive device.

Huang: Nvidia's GPUs. When they say our GPUs, people imagine little GPUs. Our GPU, one GPU is now rack scale. It's two tons, 120,000 watts, about $3 million. That's a GPU. We also sell smaller GPUs, the ones that Jeff Hinton use, that that's, like, $1,000, $500, that plugs into your PC, and you could use it for video games or AI and things

like that. But we also have bigger GPUs. And then, one gigawatt AI-factory GPU is about, you know, $50 billion.

Buhler: So tell us about these AI factories because you might have the small one might be the AI blender. But then you have the really big one, these AI factories that you went all in in 2016 and started to say, "The world is going to need AI factories." How did you get that edge, that conviction. And then-

Huang: You just gotta reason about it. Exactly. You got to reason about it. So we built the first one. It was the most expensive computer the world's ever seen, $300,000 per node. And it wasn't that successful. And so I came to the conclusion, we didn't make a big enough. And so we made a bigger one, and the second one became super successful. And, and now the question then becomes: how large do you make it and how hard do you drive computation. The reason why things are moving so fast is Nvidia's product cycles, and the way we innovate, the way we design. We're not designing a chip. We're designing an entire infrastructure all at one time. We're the only company in the world today that you can give a building some power and a blank sheet of paper, and we can create everything within it. All of the networking, all the switches, all the CPUs, all the GPUs, you know, all of the technology within that entire factory we can build and we can, and it all runs. That's the same software stack from Nvidia and because we we can integrate like that, we can also move extremely fast. So I could redesign the next year's, the redesign the next year's. And every single year, they're all software-compatible. The benefit of software compatibility is velocity. The reason why the PC was able to move so fast was because they were all Windows-compatible and, therefore, by definition, if you're compliant with the stack, you could build chips as fast as you like. And so we're now building AI factories as fast as we like at the limits of what's physically possible. And so, and because we're innovating at such incredible scale and we're co-designing, meaning we're changing algorithms are changing software, we're changing networking and CPUs and GPUs all the same time. We break out of Moore's Law's limits, which is, as you know, slowing down. And so generationally, we introduce performance levels by about 10 times. I mean, that's an incredible level of performance that we give to the market every single, every single year. The reason why we do that is we believe that just around the horizon is a problem that is so large you need a larger computer, faster computer, on the one hand. On the other hand, when we increase performance at the same power, we're decreasing your cost. And so we're driving costs down incredibly fast, which allows customers to do bigger things, which allows them to generate more revenues from the same factory. And so Nvidia's, you know, adoption today is because we are both the highest performance. We're the highest scale. And so if you want giant systems, you could do so. And we're the lowest cost. Our performance is so high. You know, for example, if

you're if your data center's one gigawatt, you're not going to get more than that. You're one gigawatt. And so if our perf per watt, our energy performance per unit of energy used is three times, your company can generate three times more revenues in that factory. That's why I call it a factory. It's not a data center. It's a factory. They're making money from it. And so these AI factories want to keep driving the scale up. They want to keep driving the revenues up. They want to keep driving the throughput up. And so that's the reason why we're innovating so fast. And it's hard to keep up with us. And it also explains why we're successful.

Buhler: Jensen, you have shifted from a component to a whole platform. That's the AI-factory concept. For an investor audience, can you break down what goes into the platform and then, also, start to talk about what's next for what the platform looks like?

Huang: Well, the, you know, they're CPUs, GPUs, network processors. There are three types of switches. There's a scale-up switch that turns one rack into a whole computer. We invented rack-scale computing. It's called "scaling up." You scale it out by taking a whole lot of these racks and connecting them together. That switch and that networking has a bunch of software on it. Software on top of all this stuff. And then you take, you create one giant system, the size of this building. And this, this building would probably be about 100MW. A gigawatt is a few thousand acres. And then you connect all these data centers together with even networking so that all of the data centers can think together. And so that's what we built together. That's what we built today. There are several reasons why infrastructure is being built so fast. And, and, there's some questions that, that, floating around about about the bubble and comparing it to the year 2000. And so just, just to compare it, during the time of 2000, internet companies- There were Hospital.com. There was Pets.com. Most of the internet companies were not profitable. And the size of the whole internet industry was about $20 or $30 billion, if you recall. And today, the the first thing you need to observe is that AI isn't just about the brand new companies, OpenAI and Anthropic and others. AI is transforming the way that hyper-scalers do work. Like, for example, search is now powered by AI recommender systems. How you see ads and news and stories are now movies generated by, recommender, by AI, user-generated content. So basically, Google's business, Amazon's business, Meta's business, hundreds of billions of dollars of revenues are all powered by AI now. Even in the absence of OpenAI and Anthropic, this entire hyper-scale industry is being powered by AI. And so the first thing to observe is that whole thing needs to go from classical CPUs with classical machine learning to now deep learning with AI. So that transition alone is hundreds of billions of dollars. Does it make sense? Absolutely. And so that's one. The second thing is that we now have this new market. This new market is called, you know, AI and and it's got a new industry and they produce AI. And so the OpenAIs, the

Anthropics the xAIs, the Geminis from Google, of course. And Meta is going to be an AI maker. And so this entire layer of AI model makers is also building AI factories. And these AIs are going to power the next generation of new opportunities. And this is where the Harveys, the OpenEvidences, the Cursors, I mean, you're right. You see all of these AI-native companies, and they're going to be connected to AI models, and they're going to they're going to go after, for the very first time in history, an industry that never was addressable. And that's the labor industry. And it's digital labor, digital, call it "agentic AI," is going is going to supplement and augment the enterprise market. So, for example, Nvidia already today we use 100% of our software engineers, 100% of our chip designers. Every single engineering today is augmented by Cursor. We use Cursor, largely, inside our company. And so we now have AIs for all of our engineers. Productivity gains. The work that we do is so much better. You also see that there's a new industry showing up. It's called "physical AI." So you have enterprise AI, you have physical AI, or augmenting labor. And so, for example, a robo-taxi is essentially a digital chauffeur. Right? And we're now going to have AIs that are going to be embodying, going to embed into anything that moves. And so, in the case of, robo-taxi, it's a steering wheel and wheels. But you're going to pick a, you know, pick and place arms. You're gonna have one arm, two arms. You know, three legs, All kinds of different, embodiments. And so these two industries represent about $100 trillion of the world's economy. And for the very first time, we have technology that's going to be able to augment that. And so that's the reason why, you know, people are so excited about the next, next wave of AI.

Buhler: So let's talk for a moment on the previous wave because you mentioned how AI has already been offering an ROI. And for the investor audience, I think the Meta example is a great case study because in Q4 2022, Apple basically removed attribution data from Meta, and you all saw hundreds of billions of dollars of market cap decline. And the Meta team said, "How are we going to fix this?" They fix that with AI powered by Nvidia GPUs.

Huang: That's right. Yeah.

Buhler: And they got their attribution back up to where it was. And that has recovered many hundreds of billions. It's over a trillion higher than it was at it's low. And that is all ROI that was powered, really, by your GPUs.

Huang: What Meta was, classically, not just Meta, but, one of the most complicated systems, software systems, is called a recommender system. And there's a couple of basic technologies. One of them is called "collaborative filtering," which is based on what I'm doing and looking at what everybody else is doing. If we have similar patterns, it would recommend maybe the same movie to me, the same next item in your grocery

list, you know, a book to me, a video to me, so on and so forth. And then the other thing is called "a content filtering," just based on who I am and my preferences. And based on what that book actually is, you might be able to recommend that book to me. And so the recommender system is the largest software ecosystem in the world. And that ecosystem is moving very significantly, very quickly to AI. And so you're going to need a, you know, a mountain of GPUs.

Buhler: And those systems were made famous by the Netflix challenge a couple decades ago. Now, Netflix, their recommendations are all powered by AI and Amazon. As you said, when you go and purchase something, a significant number is by a recommendation system.

Huang: Moving search to AI.

Buhler: Moving search to AI. All of this is being powered now.

Huang: TikTok to AI, right?

Buhler: Yeah.

Huang: Google Shorts, AI. I mean, without it- And now, all of the personalized ads are going to AI. And so just the amount of AI is just incredible. And that has nothing to do- Notice, I've just described a whole bunch of classical use cases. Quantitative trading is going to move to AI. What used to be human-engineered feature extraction is going to move towards AI.

Buhler: And I think that's actually an area that Citadel Securities has pioneered for the past 20-some years. So that's the classical AI.

Huang: Citadel said that I was a great customer. Thank you.

Buhler: So that is a classical example. And for the investor audience, talking about AI ROI, it's already there in the form of trillions of market cap. Yeah. Let's talk about what's next for spend. So 2025 estimates can be as high as $billion of AI investment in the ground. Where do we go from here? Does this become a multi-trillion dollar a year investment category?

Huang: Yes. The manufacturing, the foundry part of AI, if you will, is the model makers. They're kind of like, think of them like wafer makers. The applications of that. And one way of thinking about AI is the large language models. That's the operating system, if you will, of the modern computer. And you build applications on top of these AI models. Not just one AI model, but a system of AI models. OK. And so applications have, you know, it's going to have a collection of different AIs that it connects

together. And so the question is: What's the application space on top? The most sensible way of thinking about the application space on top, aside from all of the whatever applications we have, are going to be improved by that- We've been talking about- A simple metaphor is just digital humans. And so, a digital software engineer, right? AI coding. It's going to be a couple trillion-dollar market opportunity, probably. You know, AI digital nurses, AI accountants, AI lawyers, AI. Right. So there's AI marketeers. So we call all of that "agentic AI," and that technology is evolving very nicely. And so, for the very first time, technology is no longer just a tool used by accountants, tools used by software engineers. We're going to become digital software engineers. And I wouldn't be surprised if you you license some and you hire some. And so depending on the quality and depending on the deep expertise, and so future workforces in enterprise will be a combination of humans and digital humans, and some of them will be OpenAI-based and some of it would be Harvey-based or, you know, OpenEvidence or Curser or, you know, Replit or, you know, Lovable or some of it will be third-party and some of them you'll home-grow. And so we home-grow a lot of our own AIs because we have a lot of proprietary knowledge and data that we want to protect. And we have, we have skills in developing those AIS. Over time, more and more people will be able to cultivate their own digital AIs because it will just be easier, easier to do so. And so enterprise agentic AI, you know, obviously, augmenting the labor force is trillions of dollars of opportunity. And what's unique about AI also, versus previous software, is that AI needs to think, meaning you can't pre-compile it, put it into a binary, download it, and use it. It's gotta process all the time. And the reason why it processes, it has to take your context. It has to think about what you wanted to do and then produce an output. And so it's thinking and thinking and generating. It needs a machine. It needs computers to do that. And that's the reason why AI factories exist. And so these AI factories will be in the cloud. They might be on Prem. They'll now be all over the world. And, I, you know, it's part of the, the AI infrastructure, if you will. But there's gonna be a whole lot of thinking to produce these, these, we call them "tokens," but basically intelligence. And so that's the cognitive AI, the digital workforce, if you will. And then the second one is robotics. You know, for the very first time. Right? So let me give you a thought experiment. You know why robotics is so close. As you know, as you know, you can now prompt an AI and it can generate, you know. Prompt: "Jensen picking up a bottle, opening it and taking a sip." OK? And it would generate the video of me, right? Opening up a bottle and taking a sip. Well, if it can generate all of that, why can't it maneuver a robot to do that? And so your thought experiment would suggest that, you know, that's probably very likely. Now, if you could design a digital chauffeur that could drive a car. Why can't you have a robot, a physical robot, drive a car? And so, if a physical robot, if you can embody a physical robot to even drive a car, why can't you embody a, you know, pick-and-place arm or any type of robotics? And so, notice, we have the ability to embody almost anything we could pick

up, pick up knives and forks, and it becomes an extension of our body. And somehow, we articulated. We could pick up a baseball bat and use it as an extension of our body. And so we embody these physical extensions. Future, future AIs will be able to embody, you know, and manipulate a car, robotic arms, a humanoid robot, a surgical robot. You know, so on and so forth. And so, so I think these two, these two markets are, are within reach of, of AI. And then lastly if I just give you one example. You know, whenever you see the observation of one thing, the rest of it is just engineering. Right? And so, we know, we've now seen the, the evidence of one excellent thing, which is a rob-, a digital, an AI software coder, which is the reason why we use it so much. If you can have an AI software coder, why can't you have that AI software coder also write software to be a marketing campaign? Write software to, you know, help you solve any accounting. You know, whatever you want to do. And so almost the existence of that says the rest of it is engineering. And then, we now have robo-taxis, you know? It's an embodied robot that controls a steering wheel and wheels. And, why, that exists. Why can't you generalize that? And so the rest of it is just engineering. And so I think that's a good way to reason from first principles. How likely it is we're going to be able to have this technology proliferate across industries and society. And then the next thing that you have to reason about is, OK, so how do you scale this out? How do you deliver this intelligence to all of these different applications? Well, you need AI factories. And so.

Buhler: So let's talk a little more about robotics. You have an exceptional robotics team, one of your executives who runs robotics here today. In a previous conversation, you shared some insight about how robotics might play out. You know, is it going to be a single humanoid project? Is it going to be open source projects? How are those open source projects going to tie back? How do you think robotics will actually manifest in the physical world? And on what timeline?

Huang: Well, robo-taxis are here now. Yeah. And their ability to generalize from city to city to city is really, really getting fast. And the reason for that is because the same fundamental technology, we went through the same journey and for all the, the, the quantitative trading, the algorithmic trading, people in the room, you went from human- engineered features, machine learning to using more deep learning and, you know, embedding certain modalities and, multi-modality models to, to now, largely end to end. And the reason why, the reason why and it's multi-modal. In this journey, we became more and more generalizable, and the AI model that you use for self-driving car and the AI model that you use for a human or robot is highly similar. It's just in two different embodiments. And the reason why I know that for sure is because I can drive a car. And I can manipulate my body. It's the same intelligence. And so. And I could pick up a fork and knife and, somehow, you know, pretend like I'm a surgeon, you know, doing surgery on a steak, you know. And so, so you could notice it's the

same AI in different embodiments. And so that's, that's where AI is going. Robotics is going towards a general, more and more generalizable AIs that are multi-embodiment. It's multimodality. It's multi-embodiment. And in order to create this future, you need three things. You need the AI factory I was talking about, where you have to train the models. And you need a place where the AI can learn how to be an AI without having to come into the world right away. So we could try trillions of different iterations inside a virtual world. Well, that virtual world resembles a video game. And so the AI is basically playing a game inside a virtual world, like a video game character, and it obeys the laws of physics. And when it's done learning how to be a great video game player, because the SIM-to-real gap is extremely low because the simulator is really, really good. We call it "omniverse." That omniverse computer, then the robot can come out of that virtual world, and this world becomes one more version of the virtual worlds it played in. And it comes into the physical world. When it comes into the physical world, it needs a computer as well. So you need three computers. You need the AI computer, training computer, you need the simulation, the lab, the virtual-world computer. And then you need a computer where the robot actually operates the brain. And so Nvidia offers all three of those computers. And we work with just about every robotics company, self-driving car company, you know, robotics of different embodiments. And this is likely going to be one of the largest markets of all.

Buhler: So Nvidia touches just about everything in technology now. And as you've said in the past, you start with zero billion-dollar markets and help turn them into trillion-dollar markets. Robotics is one of the next- frontier markets. Are there any other next-frontier markets that you're particularly excited about? You mentioned healthcare a moment ago. Is that one you're passionate about? Are there others that the investors in the room should be on the lookout for?

Huang: Well, the technology, the technology needed for healthcare is really complicated. And we're making fast progress. If you can understand the meaning of words, sequences of characters, you might, you might, you might, and you could understand the meaning of structures like the virtual world. OK? Like when you, when you look at the reason why we're able to generate video is because we understand the virtual world to generate an image, a representation of the virt-, of the world. And so, if you can generate video, it must be because you understand the world. If you can generate, if you can understand worlds, is it possible you understand proteins and chemicals that have structure? And the answer is yes. And that's it. We're increasingly, getting closer to closer to understand the meaning of proteins, AlphaFold and others. We're able to understand the meaning of cells. And we recently did a partnership with Arche, and Evo is one of the first examples of a large language model that a foundation model for cell representation. So you, you can now talk to it and say, "I want

you to generate other cells of these properties." And, or, you could talk to a cell. You know, "What are what are your, your, properties, and what can you bind to? And what can, but, your metabolism, what can you activate with?" And so, you could talk to a cell like you could talk to a chatbot. And so, understanding the meaning of proteins, you know. Anyway, so there's a lot of progress there. I mean, the list goes on. I mean, the, I'm excited about the work that we're doing to bring AI into telecommunications. 5G and 6G will be revolutionized by AI. I'm excited about the collaboration we have with quantum computers so that we can, we can pull in the quantum computer schedule by about a decade by creating quantum GPU hybrid computing systems, where we do the error correction, we control the quantum computer, we do the post-processing. And so we have a new architecture called CUDA Q, which extends CUDA to quantum. And that's getting incredible adoption. And so, yeah, there's there's a whole bunch of problems we can now solve that were hard to solve before.

Buhler: Let's talk a little bit about sovereign AI. We just had Mario Draghi on the stage. He was talking about the importance of new investments in technology for the European Union, including, obviously, AI at a large scale. This revolution is materially different in that governments are highly involved both in potentially regulating but also in purchasing AI factories. Can you tell us, what do you think is the way forward? Both for sovereign AI, how come countries have their own AI systems, and also for import/exports. How we, as the United States, should be interfacing with the rest of the world with AI?

Huang: Well. No country can afford to outsource all of their nation's data. So that, and import your own intelligence back to yourself. And I just think, on first principle, that's not sensible. And, however, nobody needs to only build everything themselves. You could you could buy. You could import. But you shouldn't give up on the production of your own national intelligence. And so I think the, today the technology is rather hard. But it's getting easier and easier very, very quickly. And there's an enormous amount of open source capability. And so I would, I wouldn't give up on building your own sovereign AI. I wouldn't give up on taking the data that you have and creating your own national intelligence from it. And now countries all over the world are doing so. And so I think sovereign AI is likely, every country is likely to import some, buy some and also build some. And there's a lot of capabilities to doing that. And so we're seeing just a lot of momentum around sovereign AI. The UK's doing it. You know, I was in France. We support a company called Mistral. In the UK, there's a company called Nscale. There's a company called Nebius. In, in, in Italy, there's a, there's, several companies. In Spain, there's several companies. Germany, there's several companies. And so there's companies all over the world. And in Japan, there's companies, you

know, in Korea there's companies. And so, they're- sovereign AI's cropping up all over the world.

Buhler: Yeah. So one country that's come up a lot is China. What's the right thing for the United States in terms of exports to China of AI factories?

Huang: Well, AI is a new technology, and we have to think about, before we, yeah, we have to be thoughtful about ultimately how to regulate it. United States, of course, wants to win the AI race. And I think the policymakers all want to do the right thing, and they want America to win. However, it's important to be mindful that what is, what harms China could, oftentimes, also harm America and even worse. And so before we leap towards policies that are hurtful to other people, take a step back and maybe reflect on what are the policies that are helpful to America. And it probably is the case that you have to go back to first principles again. In the case of AI, what's most important about AI and any computing, any software industry, the developers are vitally important, as you know. And so winning developers is what creates the future platform. And we want the world to be built on American technology, you know, and Nvidia is a proud American company. And we want, we want, of course, we hope that we could create American technology that the world's built upon. Well, a lot of the AI researchers are in China. You know, China has about 50% of the world's AI researchers, incredible schools, incredible focus in AI, lots of passion around AI. And I think it's a mistake to not have those researchers build AI on American technology on first principles. I think that's a mistake. And so the question is: How do you balance winning, staying ahead? On the other hand, ensuring that the world builds on American tech stack. That's the balance. And in order to balance, you have to have nuance. And it's probably not, you know, all or nothing. And so nuance, a nuanced strategy that that, changes that, that, you know, is changing over time, has a, you know, that allows the United States to stay ahead while we continue to win researchers around the world. It's probably the right balance. And that's what I would advocate. At the moment, we are 100% out of China. And so China is 0%. We went from 95% market share to 0%. And so I can't imagine any policymaker thinking that that's a good idea. That whatever policy we implemented caused, one of, caused America to lose one of the largest markets in the world to 0%. But anyhow, and all of our forecast, if there any shareholders out there, all of our forecast, we're assuming zero for China. If anything happens in China, which I hope it will, it'll be a bonus. But it's a large market. China's the second largest computer market in the world. It is a vibrant ecosystem. I think it's a mistake for United States to not participate. And so, hopefully, we'll we'll continue to explain and inform and hold out hope for a change in policy.

Buhler: Jensen, you were at our offices recently for an AI conference we were holding, and you had some really brilliant insights into the future of AI security and the

importance of it. It's somewhat related. There are nation-state actors that might interfere with AI. There are individual users that might use AI incorrectly. What do you think the future of AI security looks like?

Huang: Well, AI security in the future is going to look a little bit like cybersecurity. It's going to require that, that we all work as a community. You probably know this. All of your cybersecurity, your chief security officers, we're all one large community. And when somebody finds a, you know, some intrusion, we share with everybody. Whenever we find a vulnerability, we share with everybody. And so, it's very likely that the future of AI security will be like cybersecurity. Second, if the marginal cost of intelligence, the marginal cost of AI goes to zero, if the marginal cost of AI goes to zero, then why wouldn't the marginal cost of security-focused? AI go also to zero? And so that's very clear. It's very likely that every AI will be surrounded by a whole bunch of cybersecurity AIs watching it. And so we're going to have lots and lots of AI protectors. Thousands of them, millions of them inside the company, outside the company. And so that's kind of the future of, the idea that that an AI has to itself be good is good. But I don't think we should rely on it. And so just like the idea that a piece of software should be properly functioning, we like, but the idea that you could have bugs or, you know, it could be a virus or whatever it is, it could be an intruder, we have to assume. And so we're going to make AI advance as safely as possible. And then, we're all going to also surround AI with a lot of security AIs.

Buhler: You shared that, really, the dynamics of the physical world are decoupled in this digital world. Where in the physical world, you might have one security person to 100 normal people, it could be inverted in an AI world. You also shared this idea that I found. mind-expanding.

Huang: For example, like cybersecurity.

Buhler: Yeah.

Huang: Yeah, we have a lot more cybersecurity agents than we have people working in the company on cybersecurity.

Buhler: You also shared this idea that in the future, we're not just going to have rendered computation, but everything is going to be generated. Can you unpack what that prediction is and what that means for Nvidia?

Huang: Well, the greatest, the best example of that. A couple of examples. Perplexity. Everything that you see on Perplexity when you ask a question is completely generated. 100% of everything you see is generated. And yet, in the past, before Perplexity, you would type in something, and it would give you a list, and you would go

and click on it, and all of that content was written by somebody or created by somebody a priori. So search is storage-based computing. It's retrieval-based computing. It's retrieving information for you to consume yourself. Perplexity, or AI, is generating. It goes and studies. It goes and reads all the content, and it generates it for you. OK, so Perplexity is a great example of the classical computer approach, we go retrieve a file and read it, to a generative approach, Perplexity, which is AI-based. Another one, which is, look at the videos that we see today. You know, Sora is, of course, Nano Banana, of course, you know, all of those pixels are generated. It's conditioned and prompted by you. You know, you might give it an initial seed of something, and say, you know, "I would like you to generate a video of Konstantine and Jensen having a fireside chat." And then, you would prompt it and say, "This fireside chat is going to talk. They're going to talk about, you know, crazy stuff."

Buhler: For those online, this is real, actually.

Huang: And then, Sora would generate it. And so. Every single pixel, every single motion, every single word is generated. So the way of computation in the future will likely be generative. And let me just give you one final idea. 100% of what you and I just went through is generated. Every question you asked me. I didn't run back to my office and retrieve something and bring it to you. "Is this what you meant, Konstantine?" And then you read it aloud for everybody to hear. That's yesterday's computer. Today's computer is, just- We're just interacting. And so we are generating everything in real time based on the context that is happening right here based on the audience, based on what's happening around the world. And so, we're generating everything in real time. That's the future computer. You know, your future computer is a CEO in front of you, or it's an artist. It's a, you know, it's a poet. It's a storyteller. And you collaborate with it to create unique content for yourself. And so, the future of computation is 100% generative. And behind it, you need an AI factory, which is the reason why I'm 100% certain we're at the beginning of this journey, and, you know, we're a few hundred billion dollars, a few, extremely small. We're only a few hundred billion dollars of infrastructure built for what likely will be trillions of dollars of infrastructure built each year. And so that's the easiest way to think through it.

Buhler: And that computing paradigm is so much more like the human mind.

Huang: Yeah. It's thinking, you know.

Buhler: So. If you're up for it, how about we generate a few lightning-round answers? OK?

Huang: OK.

Buhler: OK. Just in the last few minutes together–

Huang: I'm sure fried chicken is the answer.

Buhler: I don't know what the question is for that one. So let's jump in. What's one KPI that Wall Street underweights?

Huang: In the future of AI factories, your throughput-per-unit energy governs the revenues of your customers. It's not just about selecting a better chip. It's about deciding what your revenues are going to be. And in fact, if you go back and look at all the CSPs, the ones that chose right, saw revenue growth. And the ones that slow down subsequently chose right. And so, you could see, you could see it playing out. And people are starting to understand it. Your throughput. Token. It's called tokens. Token generation rate per unit energy of your factory is your revenues.

Buhler: The most underrated piece of Nvidia's platform.

Huang: Most people talk about CUDA, and CUDA is very important, but there's a suite of libraries that sit on top of CUDA, and I mentioned one earlier today. It's called Q DNN, and it is probably one of the most important libraries ever created in the history of humanity. The past, the previous one was called SQL, S-Q-L. And this one, Q-D-N-N. There's a few others. Cydia. cuLitho, which is going to be used for semiconductor manufacturing, lithography. We have about 350 of these libraries, and these libraries, that is Nvidia's treasure trove.

Buhler: What's one technology that you think is wildly undervalued and one that you think might be overvalued?

Huang: Undervalued. Undervalued. Undervalued. Wow. I think that the virtual world for physical AI to learn to be to a good physical AI, we call it "omniverse," is hard to understand, but it is, it is deeply undervalued and not because people use it and don't know. They just don't know they need it yet. But now, omniverse is sweeping across the robotics industry, and everybody now gets it. Once you start building robots, you'll start to realize, you know, how visionary it was that we started working on omniverse almost a decade ago. And so omniverse is really important.

Buhler: What's the book that most shaped your business and leadership philosophy?

Huang: One of my favorite books was the, you know, everybody's first calculus book. That's when you realize that math was in motion. That was a good book. All of Clay's books, Christiansen's books, and he's passed. But a good friend, all of his books were great. Al Ries' "Positioning" book, really good book, if you haven't had a chance. Of course, you know "Sapiens" always good. But those are good ones. You know,

Jeffrey's book, on "Crossing the Chasm." That's a good book. But all of Christiansen's books. Read them all.

Buhler: Favorite comfort food?

Huang: There you go, fried chicken.

Buhler: There you go. OK, we got it in. Alright. And then, last question: If you were a CIO in the audience with $10 billion to allocate toward AI in the coming years, what would you invest it into?

Huang: I would, right away, experiment with building your own AI. I mean, I just, you know, the fact of the matter is we take pride in onboarding employees and how the method by which you do so, the culture by which you bring them into, the philosophies of your company, the operating methods, the practices that makes your company what it is. The, the collection of data and knowledge that you've embodied over time that you make accessible to them. And so that is what defines a company in the past. A company of the future includes that, of course, but you need to do that for AI. You need to onboard digital. You need to onboard AI employees. There's methodology for onboarding AI employees for, we call them "fine-tuning," but basically teaching them, you know, the, the culture of, the knowledge of, the skills of, evaluation methods. And so the entire flywheel of your agenetic employee is something that you need to go and learn how to do. I tell my CIO, our company's IT department, they're going to be the HR department of agentic AI in the future. They're going to be the HR department of digital employees of the future. And those digital employees are going to work with our, of course, biological ones. And that's going to be the shape of our company in the future. And so if you get a chance to do that, I, we do that right away.

Buhler: Thank you, Jensen. Well, we heard an incredible story. Really, the story of Nvidia is one of exceptional generalization from an accelerated graphics processor to the technology that powers all of AI in the world today, from a component and the world's first GPU to all of the components in a platform in the world's AI factory. We talked about how services are the baseline for this new revolution and how robotics are in all of our future. We covered foreign policy. We even touched on fried chicken. You did it all, Jensen. Thank you so much.

Huang: Thank you. Good job. Thank you.