

NVIDIA CEO Jensen Huang's Vision for the Future

62 MIN · YOUTUBE · [HTTPS://WWW.YOUTUBE.COM/WATCH?V=7ARBJQN6QkM](https://www.youtube.com/watch?v=7ARBJQN6QkM)
<https://www.youtube.com/watch?v=7ARBJQN6QkM>

At some point, you have to believe something.

We've reinvented computing as we know it. What is the vision for what you see coming next? We asked ourselves, if it can do this, how far can it go? How do we get from the robots that we have now to the future world that you see? Cleo, everything that moves will be robotic someday and it will be soon. We invested tens of billions of dollars before it really happened. No that's very good, you did some research! But the big breakthrough I would say is when we...

That's Jensen Huang, and whether you know it or not his decisions are shaping your future. He's the CEO of NVIDIA, the company that skyrocketed over the past few years to become one of the most valuable companies in the world because they led a fundamental shift in how computers work unleashing this current explosion of what's possible with technology. "NVIDIA's done it again!" We found ourselves being one of the most important technology companies in the world and potentially ever. A huge amount of the most futuristic tech that you're hearing about in AI and robotics and gaming and self-driving cars and breakthrough medical research relies on new chips and software designed by him and his company. During the dozens of background interviews that I did to prepare for this what struck me most was how much Jensen Huang has already influenced all of our lives over the last 30 years, and how many said it's just the beginning of something even bigger. We all need to know what he's building and why and most importantly what he's trying to build next. Welcome to Huge Conversations...

Thank you so much for doing this. I'm so happy to do it. Before we dive in, I wanted to tell you how this interview is going to be a little bit different than other interviews I've seen you do recently. Okay! I'm not going to ask you any questions about - you could ask - company finances, thank you! I'm not going to ask you questions about your management style or why you don't like one-on ones. I'm not going to ask you about regulations or politics. I think all of those things are important but I think that our audience can get them well covered elsewhere. Okay.

What we do on huge if true is we make optimistic explainer videos and we've covered - I'm the worst person to be an explainer video. I think you

might be the best and I think that's what I'm really hoping that we can do together is make a joint explainer video about how can we actually use technology to make the future better. Yeah. And we do it because we believe that when people see those better futures, they help build them. So the people that you're going to be talking to are awesome. They are optimists who want to build those better futures but because we cover so many different topics, we've covered supersonic planes and quantum computers and particle colliders, it means that millions of people come into every episode without any prior knowledge whatsoever. You might be talking to an expert in their field who doesn't know the difference between a CPU and a GPU or a 12-year-old who might grow up one day to be you but is just starting to learn. For my part, I've now been preparing for this interview for several months, including doing background conversations with many members of your team but I'm not an engineer. So my goal is to help that audience see the future that you see so I'm going to ask about three areas: The first is, how did we get here? What were the key insights that led to this big fundamental shift in computing that we're in now? The second is, what's actually happening right now? How did those insights lead to the world that we're now living in that seems like so much is going on all at once? And the third is, what is the vision for what you see coming next? In order to talk about this big moment we're in with AI I think we need to go back to video games in the '90s. At the time I know game developers wanted to create more realistic looking graphics but the hardware couldn't keep up with all of that necessary math. NVIDIA came up with a solution that would change not just games but computing itself. Could you take us back there and explain what was happening and what were the insights that led you and the NVIDIA team to create the first modern GPU? So in the early '90s when we first started the company we observed that in a software program inside it there are just a few lines of code, maybe 10% of the code, does 99% % of the processing and that 99% of the processing could be done in parallel. However the other 90% of the code has to be done sequentially. It turns out that the proper computer the perfect computer is one that could do sequential processing and parallel processing not just one or the other. That was the big observation and we set out to build a company to solve computer problems that normal computers can't. And that's really the beginning of NVIDIA.

My favorite visual of why a CPU versus a GPU really matters so much is a 15-year-old video on the NVIDIA YouTube channel where the Mythbusters, they use a little robot shooting paintballs one by one to show solving problems one at a time or sequential processing on a CPU, but then they roll out this huge robot that shoots all of the paintballs at once doing smaller problems all at the same time or parallel processing on a GPU.

"3... 2... 1..." So Nvidia unlocks all of this new power for video games. Why gaming first? The video games requires parallel processing for processing 3D graphics and we chose video games because, one, we loved the application, it's a simulation of virtual worlds and who doesn't want to go to virtual worlds and we had the good observation that video games has potential to be the largest market for for entertainment ever. And it turned out to be true. And having it being a large market is important because the technology is complicated

and if we had a large market, our R&D budget could be large, we could create new technology. And that flywheel between technology and market and greater technology was really the flywheel that got NVIDIA to become one of the most important technology companies in the world. It was all because of video games. I've heard you say that GPUs were a time machine? Yeah. Could you tell me more about what you meant by that? A GPU is like a time machine because it lets you see the future sooner. One of the most amazing things anybody's ever said to me was a quantum chemistry scientist. He said, Jensen, because of NVIDIA's work, I can do my life's work in my lifetime. That's time travel. He was able to do something that was beyond his lifetime within his lifetime and this is because we make applications run so much faster and you get to see the future. And so when you're doing weather prediction for example, you're seeing the future when you're doing a simulation a virtual city with virtual traffic and we're simulating our self-driving car through that virtual city, we're doing time travel. So parallel processing takes off in gaming and it's allowing us to create worlds in computers that we never could have before and and gaming is sort of this this first incredible case of parallel processing unlocking a lot more power and then as you said people begin to use that power across many different industries. The case of the of the quantum chemistry researcher, when I've heard you tell that story it's that he was running molecular simulations in a way where it was much faster to run in parallel on NVIDIA GPUs even then than it was to run them on the supercomputer with the CPU that he had been using before. Yeah that's true. So oh my god it's revolutionizing all of these other industries as well, it's beginning to change how we see what's possible with computers and my understanding is that in the early 2000s you see this and you realize that actually doing that is a little bit difficult because what that researcher had to do is he had to sort of trick the GPUs into thinking that his problem was a graphics problem. That's exactly right, no that's very good, you did some research. So you create a way to make that a lot easier. That's right Specifically it's a platform called CUDA which lets programmers tell the GPU what to do using programming languages that they already know like C and that's a big deal because it gives way more people easier access to all of this computing power. Could you explain what the vision was that led you to create CUDA? Partly researchers discovering it, partly internal inspiration and and partly solving a problem. And you know a lot of interesting interesting ideas come out of that soup. You know some of it is aspiration and inspiration, some of it is just desperation you know. And so in the case of CUDA is very much this the same way and probably the first external ideas of using our GPUs for parallel processing emerged out of some interesting work in medical imaging a couple of researchers at Mass General were using it to do CT reconstruction. They were using our graphics processors for that reason and it inspired us.

Meanwhile the problem that we're trying to solve inside our company has to do with the fact that when you're trying to create these virtual worlds for video games, you would like it to be beautiful but also dynamic. Water should flow like water and explosions should be like explosions. So there's particle physics you want to do, fluid dynamics you want to do and that is much harder to do if your pipeline is only able to do computer graphics. And so we have a natural reason to want to do it in the market that we were serving. So researchers were also horsing around with using our GPUs for general purpose uh acceleration and

and so there there are multiple multiple factors that were coming together in that soup, we just when the time came and we decided to do something proper and created a CUDA as a result of that. Fundamentally the reason why I was certain that CUDA was going to be successful and we put the whole company behind it was because fundamentally our GPU was going to be the highest volume parallel processors built in the world because the market of video games was so large and so this architecture has a good chance of reaching many people. It has seemed to me like creating CUDA was this incredibly optimistic "huge if true" thing to do where you were saying, if we create a way for many more people to use much more computing power, they might create incredible things. And then of course it came true. They did.

In 2012, a group of three researchers submits an entry to a famous competition where the goal is to create computer systems that could recognize images and label them with categories. And their entry just crushes the competition. It gets way fewer answers wrong. It was incredible. It blows everyone away. It's called AlexNet, and it's a kind of AI called the neural network. My understanding is one reason it was so good is that they used a huge amount of data to train that system and they did it on NVIDIA GPUs. All of a sudden, GPUs weren't just a way to make computers faster and more efficient they're becoming the engines of a whole new way of computing. We're moving from instructing computers with step-by-step directions to training computers to learn by showing them a huge number of examples. This moment in 2012 really kicked off this truly seismic shift that we're all seeing with AI right now. Could you describe what that moment was like from your perspective and what did you see it would mean for all of our futures? When you create something new like CUDA, if you build it, they might not come. And that's always the cynic's perspective however the optimist's perspective would say, but if you don't build it, they can't come. And that's usually how we look at the world. You know we have to reason about intuitively why this would be very useful. And in fact, in 2012 Ilya Sutskever, and Alex Krizhevsky and Geoff Hinton in the University of Toronto the lab that they were at they reached out to a GeForce GTX 580 because they learned about CUDA and that CUDA might be able to be used as a parallel processor for training AlexNet and so our inspiration that GeForce could be the vehicle to bring out this parallel architecture into the world and that researchers would somehow find it someday was a good was a good strategy. It was a strategy based on hope, but it was also reasoned hope. The thing that really caught our attention was simultaneously we were trying to solve the computer vision problem inside the company and we were trying to get CUDA to be a good computer vision processor and we were frustrated by a whole bunch of early developments internally with respect to our computer vision effort and getting CUDA to be able to do it. And all of a sudden we saw AlexNet, this new algorithm that is completely different than computer vision algorithms before it, take a giant leap in terms of capability for computer vision. And when we saw that it was partly out of interest but partly because we were struggling with something ourselves. And so we were highly interested to want to see it work. And so when we when we looked at AlexNet we were inspired by that. But the big breakthrough I would say is when we when we saw AlexNet, we asked ourselves you know, how far can AlexNet go? If it can do this with computer vision, how far can it go? And if it if it could go to the limits of what we think it could go, the type of problems it could solve, what would it mean for

the computer industry? And what would it mean for the computer architecture? And we were, we rightfully reasoned that if machine learning, if the deep learning architecture can scale, the vast majority of machine learning problems could be represented with deep neural networks. And the type of problems we could solve with machine learning is so vast that it has the potential of reshaping the computer industry altogether, which prompted us to re-engineer the entire computing stack which is where DGX came from and this little baby DGX sitting here, all of this came from that observation that we ought to reinvent the entire computing stack layer by layer by layer. You know computers, after 65 years since IBM System 360 introduced modern general purpose computing, we've reinvented computing as we know it. To think about this as a whole story, so parallel processing reinvents modern gaming and revolutionizes an entire industry then that way of computing that parallel processing begins to be used across different industries. You invest in that by building CUDA and then CUDA and the use of GPUs allows for a a step change in neural networks and machine learning and begins a sort of revolution that we're now seeing only increase in importance today... All of a sudden computer vision is solved. All of a sudden speech recognition is solved. All of a sudden language understanding is solved. These incredible problems associated with intelligence one by one by one by one where we had no solutions for in past, desperate desire to have solutions for, all of a sudden one after another get solved you know every couple of years. It's incredible.

Yeah so you're seeing that, in 2012 you're looking ahead and believing that that's the future that you're going to be living in now, and you're making bets that get you there, really big bets that have very high stakes. And then my perception as a lay person is that it takes a pretty long time to get there. You make these bets - 8 years, 10 years - so my question is: If AlexNet that happened in 2012 and this audience is probably seeing and hearing so much more about AI and NVIDIA specifically 10 years later, why did it take a decade and also because you had placed those bets, what did the middle of that decade feel like for you? Wow that's a good question. It probably felt like today. You know to me, there's always some problem and then there's some reason to be to be impatient. There's always some reason to be happy about where you are and there's always many reasons to carry on. And so I think as I was reflecting a second ago, that sounds like this morning! So but I would say that in all things that we pursue, first you have to have core beliefs. You have to reason from your best principles and ideally you're reasoning from it from principles of either physics or deep understanding of the industry or deep understanding of the science, wherever you're reasoning from, you reason from first principles. And at some point you have to believe something. And if those principles don't change and the assumptions don't change, then you, there's no reason to change your core beliefs. And then along the way there's always some evidence of you know of success and and that you're leading in the right direction and sometimes you know you go a long time without evidence of success and you might have to course correct a little but the evidence comes. And if you feel like you're going in the right direction, we just keep on going.

The question of why did we stay so committed for

so long, the answer is actually the opposite: There was no reason to not be committed because we are, we believed it. And I've believed in NVIDIA for 30 plus years and I'm still here working every single day. There's no fundamental reason for me to change my belief system and I fundamentally believe that the work we're doing in revolutionizing computing is as true today, even more true today than it was before. And so we'll stick with it you know until otherwise. There's of course very difficult times along the way. You know when you're investing in something and nobody else believes in it and cost a lot of money and you know maybe investors or or others would rather you just keep the profit or you know whatever it is improve the share price or whatever it is.

But you have to believe in your future. You have to invest in yourself. And we believe this so deeply that we invested you know tens of billions of dollars before it really happened. And yeah it was, it was 10 long years. But it was fun along the way. How would you summarize those core beliefs? What is it that you believe about the way computers should work and what they can do for us that keeps you not only coming through that decade but also doing what you're doing now, making bets I'm sure you're making for the next few decades? The first core belief was our first discussion, was about accelerated computing. Parallel computing versus general purpose computing. We would add two of those processors together and we would do accelerated computing. And I continue to believe that today. The second was the recognition that these deep learning networks, these DNNs, that came to the public during 2012, these deep neural networks have the ability to learn patterns and relationships from a whole bunch of different types of data. And that it can learn more and more nuanced features if it could be larger and larger. And it's easier to make them larger and larger, make them deeper and deeper or wider and wider, and so the scalability of the architecture is empirically true. The fact that model size and the data size being larger and larger, you can learn more knowledge is also true, empirically true. And so if that's the case, you could you know, what what are the limits? There not, unless there's a physical limit or an architectural limit or mathematical limit and it was never found, and so we believe that you could scale it. Then the question, the only other question is: What can you learn from data? What can you learn from experience? Data is basically digital versions of human experience. And so what can you learn? You obviously can learn object recognition from images. You can learn speech from just listening to sound. You can learn even languages and vocabulary and syntax and grammar and all just by studying a whole bunch of letters and words. So we've now demonstrated that AI or deep learning has the ability to learn almost any modality of data and it can translate to any modality of data. And so what does that mean?

You can go from text to text, right, summarize a paragraph. You can go from text to text, translate from language to language. You can go from text to images, that's image generation. You can go from images to text, that's captioning. You can even go from amino acid sequences to protein structures. In the future, you'll go from protein to words: "What does this protein do?" or "Give me an example of a protein that has these properties." You know identifying a drug target. And so you could just see that all of these problems are around

the corner to be solved. You can go from words to video, why can't you go from words to action tokens for a robot? You know from the computer's perspective how is it any different? And so it opened up this universe of opportunities and universe of problems that we can go solve. And that gets us quite excited. It feels like we are on the cusp of this truly enormous change.

When I think about the next 10 years, unlike the last 10 years, I know we've gone through a lot of change already but I don't think I can predict anymore how I will be using the technology that is currently being developed. That's exactly right. I think the last 10, the reason why you feel that way is, the last 10 years was really about the science of AI. The next 10 years we're going to have plenty of science of AI but the next 10 years is going to be the application science of AI. The fundamental science versus the application science. And so the the applied research, the application side of AI now becomes: How can I apply AI to digital biology?

How can I apply AI to climate technology? How can I apply AI to agriculture, to fishery, to robotics, to transportation, optimizing logistics? How can I apply AI to you know teaching? How do I apply AI to you know podcasting right? I'd love to choose a couple of those to help people see how this fundamental change in computing that we've been talking about is actually going to change their experience of their lives, how they're actually going to use technology that is based on everything we just talked about. One of the things that I've now heard you talk a lot about and I have a particular interest in is physical AI. Or in other words, robots - "my friends!" - meaning humanoid robots but also robots like self-driving cars and smart buildings or autonomous warehouses or autonomous lawnmowers or more. From what I understand, we might be about to see a huge leap in what all of these robots are capable of because we're changing how we train them. Up until recently you've either had to train your robot in the real world where it could get damaged or wear down or you could get data from fairly limited sources like humans in motion capture suits. But that means that robots aren't getting as many examples as they'd need to learn more quickly.

But now we're starting to train robots in digital worlds, which means way more repetitions a day, way more conditions, learning way faster. So we could be in a big bang moment for robots right now and NVIDIA is building tools to make that happen. You have Omniverse and my understanding is this is 3D worlds that help train robotic systems so that they don't need to train in the physical world.

That's exactly right. You just just announced Cosmos which is ways to make that 3D universe much more realistic. So you can get all kinds of different, if we're training something on this table, many different kinds of lighting on the table, many different times of day, many different you know experiences for the robot to go through so that it can get even more out of Omniverse. As a kid who grew up loving Data on Star Trek, Isaac Asimov's books and just dreaming about a future with robots, how do we get from the robots that we have now to the future world that you see of robotics?

Yeah let me use language models maybe ChatGPT as a reference for understanding Omniverse and Cosmos. So first of all when ChatGPT first came out it, it was extraordinary and it has the ability to do to basically from your prompt, generate text. However, as amazing as it was, it has the tendency to hallucinate if it goes on too long or if it pontificates about a topic it you know is not informed about, it'll still do a good job generating plausible answers.

It just wasn't grounded in the truth. And so people called it hallucination. And so the next generation shortly it was, it had the ability to be conditioned by context, so you could upload your PDF and now it's grounded by the PDF. The PDF becomes the ground truth. It could be it could actually look up search and then the search becomes its ground truth. And between that it could reason about what is how to produce the answer that you're asking for. And so the first part is a generative AI and the second part is ground truth. Okay and so now let's come into the the physical world. The world model, we need a foundation model just like we need ChatGPT had a core foundation model that was the breakthrough in order for robotics to to be smart about the physical world. It has to understand things like gravity, friction, inertia, geometric and spatial awareness. It has to uh understand that an object is sitting there even when I looked away when I come back it's still sitting there, object permanence. It has to understand cause and effect. If I tip it, it'll fall over. And so these kind of physical common sense if you will has to be captured or encoded into a world foundation model so that the AI has world common sense. Okay and so we have to go, somebody has to go create that, and that's what we did with Cosmos. We created a world language model. Just like ChatGPT was a language model, this is a world model. The second thing we have to go do is we have to do the same thing that we did with PDFs and context and grounding it with ground truth. And so the way we augment Cosmos with ground truth is with physical simulations, because Omniverse uses physics simulation which is based on principled solvers. The mathematics is Newtonian physics is the, right, it's the math we know, all of the the fundamental laws of physics we've understood for a very long time. And it's encoded into, captured into Omniverse. That's why Omniverse is a simulator. And using the simulator to ground or to condition Cosmos, we can now generate an infinite number of stories of the future. And they're grounded on physical truth. Just like between PDF or search plus ChatGPT, we can generate an infinite amount of interesting things, answer a whole bunch of interesting questions. The combination of Omniverse plus Cosmos, you could do that for the physical world. So to illustrate this for the audience, if you had a robot in a factory and you wanted to make it learn every route that it could take, instead of manually going through all of those routes, which could take days and could be a lot of wear and tear on the robot, we're now able to simulate all of them digitally in a fraction of the time and in many different situations that the robot might face - it's dark, it's blocked it's etc - so the robot is now learning much much faster. It seems to me like the future might look very different than today. If you play this out 10 years, how do you see people actually interacting with this technology in the near future? Cleo, everything that moves will be robotic someday and it will be soon. You know the the idea that you'll be pushing around a lawn mower is already kind of silly. You know maybe people do it because because it's fun but but there's no need to. And every car is

going to be robotic. Humanoid robots, the technology necessary to make it possible, is just around the corner. And so everything that moves will be robotic and they'll learn how to be a robot in Omniverse Cosmos and we'll generate all these plausible, physically plausible futures and the robots will learn from them and then they'll come into the physical world and you know it's exactly the same. A future where you're just surrounded by robots is for certain. And I'm just excited about having my own R2-D2.

And of course R2-D2 wouldn't be quite the can that it is and roll around. It'll be you know R2-D2 yeah, it'll probably be a different physical embodiment, but it's always R2. You know so my R2 is going to go around with me. Sometimes it's in my smart glasses, sometimes it's in my phone, sometimes it's in my PC. It's in my car. So R2 is with me all the time including you know when I get home you know where I left a physical version of R2. And you know whatever that version happens to be you know, we'll interact with R2. And so I think the idea that we'll have our own R2-D2 for our entire life and it grows up with us, that's a certainty now yeah. I think a lot of news media when they talk about futures like this they focus on what could go wrong. And that makes sense. There is a lot that could go wrong. We should talk about what could go wrong so we could keep it from going wrong. Yeah that's the approach that we like to take on the show is, what are the big challenges so that we can overcome them? Yeah. What buckets do you think about when you're worrying about this future? Well there's a whole bunch of the stuff that everybody talks about: Bias or toxicity or just hallucination. You know speaking with great confidence about something it knows nothing about and as a result we rely on that information. Generating, that's a version of generating fake information, fake news or fake images or whatever it is. Of course impersonation.

It does such a good job pretending to be a human, it could be it could do an incredibly good job pretending to be a specific human. And so the spectrum of areas we have to be concerned about is fairly clear and there's a lot of people who are working on it. There's some of the stuff, some of the stuff related to AI safety requires deep research and deep engineering and that's simply, it wants to do the right thing it just didn't perform it right and as a result hurt somebody. You know for example self-driving car that wants to drive nicely and drive properly and just somehow the sensor broke down or it didn't detect something. Or you know made it too aggressive turn or whatever it is. It did it poorly. It did it wrongly. And so that's a whole bunch of engineering that has to be done to make sure that AI safety is upheld by making sure that the product functions properly.

And then and then lastly you know whatever what happens if the system, the AI wants to do a good job but the system failed. Meaning the AI wanted to stop something from happening and it turned out just when it wanted to do it, the machine broke down. And so this is no different than a flight computer inside a plane having three versions of them and then so there's triple redundancy inside the system inside autopilots and then you have two pilots and then you have air traffic control

and then you have other pilots watching out for these pilots. And so that the AI safety systems has to be architected as a community such that such that these AIs one, work, function properly. When they don't function properly, they don't put people in harm's way. And that they're sufficient safety and security systems all around them to make sure that we keep AI safe. And so there's this spectrum of conversation is gigantic and and you know we have to take the parts, take the parts apart and and build them as engineers. One of the incredible things about this moment that we're in right now is that we no longer have a lot of the technological limits that we had in a world of CPUs and sequential processing. And we've unlocked not only a new way to do computing and and but also a way to continue to improve. Parallel processing has a a different kind of physics to it than the improvements that we were able to make on CPUs. I'm curious, what are the scientific or technological limitations that we face now in the current world that you're thinking a lot about? Well everything in the end is about how much work you can get done within the limitations of the energy that you have. And so that's a physical limit and the laws of physics about transporting information and transporting bits, flipping bits and transporting bits, at the end of the day the energy it takes to do that limits what we can get done. And the amount of energy that we have limits what we can get done. We're far from having any fundamental limits that keep us from advancing. In the meantime, we seek to build better and more energy efficient computers. This little computer, the the big version of it was \$250,000 - Pick up? - Yeah Yeah that's little baby DIGITS yeah. This is an AI supercomputer. The version that I delivered, this is just a prototype so it's a mockup. The very first version was DGX 1, I delivered to Open AI in 2016 and that was \$250,000. 10,000 times more power, more energy necessary than this version and this version has six times more performance. I know, it's incredible. We're in a whole in the world. And it's only since 2016 and so eight years later we've increased the energy efficiency of computing by 10,000 times. And imagine if we became 10,000 times more energy efficient or if a car was 10,000 times more energy efficient or electric light bulb was 10,000 times more energy efficient. Our light bulb would be right now instead of 100 Watts, 10,000 times less producing the same illumination. Yeah and so the energy efficiency of computing particularly for AI computing that we've been working on has advanced incredibly and that's essential because we want to create you know more intelligent systems and and we want to use more computation to be smarter and so energy efficiency to do the work is our number one priority. When I was preparing for this interview, I spoke to a lot of my engineering friends and this is a question that they really wanted me to ask. So you're really speaking to your people here. You've shown a value of increasing accessibility and abstraction, with CUDA and allowing more people to use more computing power in all kinds of other ways. As applications of technology get more specific, I'm thinking of transformers in AI for example... For the audience, a transformer is a very popular more recent structure of AI that's now used in a huge number of the tools that you've seen. The reason that they're popular is because transformers are structured in a way that helps them pay "attention" to key bits of information and give much better results. You could build chips that are perfectly suited for just one kind of AI model, but if you do that then you're making them less able to do other things. So as these specific structures or architectures of AI get more popular, my understanding is there's a debate between how much you place these bets on "burning them into the chip" or designing hardware that is very specific

to a certain task versus staying more general and so my question is, how do you make those bets? How do you think about whether the solution is a car that could go anywhere or it's really optimizing a train to go from A to B? You're making bets with huge stakes and I'm curious how you think about that. Yeah and that now comes back to exactly your question, what are your core beliefs? And the question, the core belief either one, that transformer is the last AI algorithm, AI architecture that any researcher will ever discover again, or that transformers is a stepping stone towards evolutions of transformers that are uh barely recognizable as a transformer years from now. And we believe the latter. And the reason for that is because you just have to go back in history and ask yourself, in the world of computer algorithms, in the world of software, in the world of engineering and innovation, has one idea stayed along that long? And the answer is no. And so that's kind of the beauty, that's in fact the essential beauty of a computer that it's able to do something today that no one even imagined possible 10 years ago. And if you would have, if you would have turned that computer 10 years ago into a microwave, then why would the applications keep coming? And so we believe, we believe in the richness of innovation and the richness of invention and we want to create an architecture that let inventors and innovators and software programmers and AI researchers swim in the soup and come up with some amazing ideas. Look at transformers. The fundamental characteristic of a transformer is this idea called "attention mechanism" and it basically says the transformer is going to understand the meaning and the relevance of every single word with every other word. So if you had 10 words, it has to figure out the relationship across 10 of them. But if you have a 100,000 words or if your context is now as large as, read a PDF and that read a whole bunch of PDFs, and the context window is now like a million tokens, the processing all of it across all of it is just impossible. And so the way you solve that problem is there all kinds of new ideas, flash attention or hierarchical attention or you know all the, wave attention I just read about the other day. The number of different types of attention mechanisms that have been invented since the transformer is quite extraordinary.

And so I think that that's going to continue and we believe it's going to continue and that that computer science hasn't ended and that AI research have not all given up and we haven't given up anyhow and that having a computer that enables the flexibility of of research and innovation and new ideas is fundamentally the most important thing. One of the things that I am just so curious about, you design the chips. There are companies that assemble the chips. There are companies that design hardware to make it possible to work at nanometer scale. When you're designing tools like this, how do you think about design in the context of what's physically possible right now to make? What are the things that you're thinking about with sort of pushing that limit today? The way we do it is even though even though we have things made like for example our chips are made by TSMC. Even though we have them made by TSMC, we assume that we need to have the deep expertise that TSMC has. And so we have people in our company who are incredibly good at semiconductive physics so that we have a feeling for, we have an intuition for, what are the limits of what today's semiconductor physics can do. And then we work very closely with them to discover the limits because we're trying to push the limits and so we discover the limits together.

Now we do the same thing in system engineering and cooling systems. It turns out plumbing is really important to us because of liquid cooling. And maybe fans are really important to us because of air cooling and we're trying to design these fans in a way almost like you know they're aerodynamically sound so that we could pass the highest volume of air, make the least amount of noise. So we have aerodynamics engineers in our company. And so even though even though we don't make 'em, we design them and we have to deep expertise of knowing how to have them made. And and from that we try to push the limits. One of the themes of this conversation is that you are a person who makes big bets on the future and time and time again you've been right about those bets. We've talked about GPUs, we've talked about CUDA, we've talked about bets you've made in AI - self-driving cars, and we're going to be right on robotics and - this is my question. What are the bets you're making now? the latest bet we just described at the CES and I'm very very proud of it and I'm very excited about it is the fusion of Omniverse and Cosmos so that we have this new type of generative world generation system, this multiverse generation system. I think that's going to be profoundly important in the future of robotics and physical systems.

Of course the work that we're doing with human robots, developing the tooling systems and the training systems and the human demonstration systems and all of this stuff that that you've already mentioned, we're just seeing the beginnings of that work and I think the next 5 years are going to be very interesting in the world of human robotics. Of course the work that we're doing in digital biology so that we can understand the language of molecules and understand the language of cells and just as we understand the language of physics and the physical world we'd like to understand the language of the human body and understand the language of biology. And so if we can learn that, and we can predict it. Then all of a sudden our ability to have a digital twin of the human is plausible. And so I'm very excited about that work. I love the work that we're doing in climate science and be able to, from weather predictions, understand and predict the high resolution regional climates, the weather patterns within a kilometer above your head. That we can somehow predict that with great accuracy, its implications is really quite profound. And so the number of things that we're working on is really cool. You know we we're fortunate that we've created this this instrument that is a time machine and we need time machines in all of these areas that we just talked about so that we can see the future. And if we could see the future and we can predict the future then we have a better chance of making that future the best version of it. And that's the reason why scientists want to predict the future. That's the reason why, that's the reason why we try to predict the future and everything that we try to design so that we can optimize for the best version. So if someone is watching this and maybe they came into this video knowing that NVIDIA is an incredibly important company but not fully understanding why or how it might affect their life and they're now hopefully better understanding a big shift that we've gone through over the last few decades in computing, this very exciting, very sort of strange moment that we're in right now, where we're sort of on the precipice of so many different things. If they would like to be able to look into the future a little bit, how would you advise them to prepare or to think about this moment that they're in personally with respect to how these tools

are actually going to affect them? Well there are several ways to reason about the future that we're creating. One way to reason about it is, suppose the work that you do continues to be important but the effort by which you do it went from you know being a week long to almost instantaneous. You know that the effort of drudgery basically goes to zero. What is the implication of that? This is, this is very similar to what would change if all of a sudden we had highways in this country?

And that kind of happened you know in the last Industrial Revolution, all of a sudden we have interstate highways and when you have interstate highways what happens? Well you know suburbs start to be created and all of a sudden you know distribution of goods from east to west is no longer a concern and all of a sudden gas stations start cropping up on highways and fast food restaurants show up and you know someone, some motels show up because people you know traveling across the state, across the country and just wanted to stay somewhere for a few hours or overnight, and so all of a sudden new economies and new capabilities, new economies.

What would happen if a video conference made it possible for us to see each other without having to travel anymore? All of a sudden it's actually okay to work further away from home and from work, work and live further away. And so you ask yourself kind of these questions. You know what would happen if I have a software programmer with me all the time and whatever it is I can dream up, the software programmer could write for me. You know what would, what would happen if I just had a seed of an idea and I rough it out and all of a sudden a you know a prototype of a production was put in front of me? And what how would that change my life and how would that change my opportunity? And you know what does it free me to be able to do and and so on so forth. And so I think that the next the next decade intelligence, not for everything but for some things, would basically become superhuman. But I can tell you exactly what that feels like. I'm surrounded by superhuman people, super intelligence from my perspective because they're the best in the world at what they do and they do what they do way better than I can do it. and I'm surrounded by thousands of them and yet what it it never one day caused me to think all of a sudden I'm no longer necessary. It actually empowers me and gives me the confidence to go tackle more and more ambitious things. And so suppose, suppose now everybody is surrounded by these super AIs that are very good at specific things or good at some of the things. What would that make you feel? Well it's going to empower you, it's going to make you feel confident and I'm pretty sure you probably use ChatGPT and AI and I feel more empowered today, more confident to learn something today. The knowledge of almost any particular field, the barriers to that understanding, it has been reduced and I have a personal tutor with me all of the time. And so I think that that feeling should be universal. If there's one thing that I would encourage everybody to do is to go get yourself an AI tutor right away. And that AI tutor could of course just teach your things, anything you like, help you program, help you write, help you analyze, help you think, help you reason, you know all of those things is going to really make you feel empowered and and I think

that going to be our future. We're going to become, we're going to become super humans, not because we have super, we're going to become super humans because we have super AIs. Could you tell us a little bit about each of these objects?

This is a new GeForce graphics card and yes and this is the RTX 50 Series. It is essentially a supercomputer that you put into your PC and we use it for gaming, of course people today use it for design and creative arts and it does amazing AI. The real breakthrough here and this is this is truly an amazing thing, GeForce enabled AI and it enabled Geoff Hinton, Ilya Sutskever, Alex Krizhevsky to be able to train AlexNet. We discovered AI and we advanced AI then AI came back to GeForce to help computer graphics. And so here's the amazing thing: Out of 8 million pixels or so in a 4K display we are computing, we're processing only 500,000 of them. The rest of them we use AI to predict. The AI guessed it and yet the image is perfect. We inform it by the 500,000 pixels that we computed and we ray traced every single one and it's all beautiful. It's perfect. And then we tell the AI, if these are the 500,000 perfect pixels in this screen, what are the other 8 million? And it goes it fills in the rest of the screen and it's perfect. And if you only have to do fewer pixels, are you able to invest more in doing that because you have fewer to do so then the quality is better so the extrapolation that the AI does... Exactly. Because whatever computing, whatever attention you have, whatever resources you have, you can place it into 500,000 pixels. Now this is a perfect example of why AI is going to make us all superhuman, because all of the other things that it can do, it'll do for us, allows us to take our time and energy and focus it on the really really valuable things that we do. And so we'll take our own resource which is you know energy intensive, attention intensive, and we'll dedicated to the few 100,000 pixels and use AI to superres, upres it you know to everything else. And so this this graphics card is now powered mostly by AI and the computer graphics technology inside is incredible as well. And then this next one, as I mentioned earlier, in 2016 I built the first one for AI researchers and we delivered the first one to Open AI and Elon was there to receive it and this version I built a mini mini version and the reason for that is because AI has now gone from AI researchers to every engineer, every student, every AI scientist. And AI is going to be everywhere. And so instead of these \$250,000 versions we're going to make these \$3,000 versions and schools can have them, you know students can have them, and you set it next to your PC or Mac and all of a sudden you have your own AI supercomputer. And you could develop and build AIs. Build your own AI, build your own R2-D2. What do you feel like is important for this audience to know that I haven't asked? One of the most important things I would advise is for example if I were a student today the first thing I would do is to learn AI. How do I learn to interact with ChatGPT, how do I learn to interact with Gemini Pro, and how do I learn to interact with Grok? Learning how to interact with with AI is not unlike being someone who is really good at asking questions. You're incredibly good at asking questions and and prompting AI is very very similar. You can't just randomly ask a bunch of questions and so asking an AI to be assistant to you requires some expertise and artistry and how to prompt it. And so if I were, if I were a student today, irrespective whether it's for for math or for science or chemistry or biology or doesn't matter what field of science I'm going to go into or what profession, I'm going to ask myself, how can I use AI to do my job better? If I want to be a lawyer, how can I use

AI to be a better lawyer? If I want to be a better doctor, how can I use AI to be a better doctor? If I want to be a chemist, how do I use AI to be a better chemist? If I want to be a biologist, I how do I use AI to be a better biologist? That question should be persistent across everybody. And just as my generation grew up as the first generation that has to ask ourselves, how can we use computers to do our jobs better? Yeah the generation before us had no computers, my generation was the first generation that had to ask the question, how do I use computers to do my job better? Remember I came into the industry before Windows 95 right, 1984 there were no computers in offices. And after that, shortly after that, computers started to emerge and so we had to ask ourselves how do we use computers to do our jobs better? The next generation doesn't have to ask that question but it has to ask obviously next question, how can I use AI to do my job better? That is start and finish I think for everybody. It's a really exciting and scary and therefore worthwhile question I think for everyone. I think it's going to be incredibly fun. AI is obviously a word that people are just learning now but it's just you know, it's made your computer so much more accessible. It is easier to prompt ChatGPT to ask it anything you like than to go do the research yourself. And so we've lowered a barrier of understanding, we've lowered a barrier of knowledge, we've lowered a barrier of intelligence, and and everybody really had to just go try it. You know the thing that's really really crazy is if I put a computer in front of somebody and they've never used a computer there is no chance they're going to learn that computer in a day. There's just no chance. Somebody really has to show it to you and yet with ChatGPT if you don't know how to use it, all you have to do is type in "I don't know how to use ChatGPT, tell me," and it would come back and give you some examples and so that's the amazing thing. You know the amazing thing about intelligence is it'll help you along the way and make you uh superhuman you know along the way. All right I have one more question if you have a second. This is not something that I planned to ask you but on the way here, I'm a little bit afraid of planes, which is not my most reasonable quality, and the flight here was a little bit bumpy mhm very bumpy and I'm sitting there and it's moving and I'm thinking about what they're going to say at my funeral and after - She asked good questions, that's what the tombstone's going to say - I hope so! Yeah. And after I loved my husband and my friends and my family, the thing that I hoped that they would talk about was optimism. I hope that they would recognize what I'm trying to do here. And I'm very curious for you, you've you've been doing this a long time, it feels like there's so much that you've described in this vision ahead, what would the theme be that you would want people to say about what you're trying to do?

Very simply, they made an extraordinary impact.

I think that we're fortunate because of some core beliefs a long time ago and sticking with those core beliefs and building upon them we found ourselves today being one of the most, one of the many most important and consequential technology companies in the world and potentially ever. And so we take that responsibility very seriously.

We work hard to make sure that the capabilities that we've created are available to large companies as well as individual researchers and developers, across every field of science no matter profitable or not, big or small, famous or otherwise.

And it's because of this understanding of the consequential work that we're doing and the

potential impact it has on so many people that we want to make make this capability as pervasively as possible and I do think that when we look back in a few years, and I do hope that what the next generation realized is as they, well first of all they're going to know us because of all the you know gaming technology we create. I do think that we'll look back and the whole field of digital biology and life sciences has been transformed. Our whole understanding of of material sciences has completely been revolutionized. That robots are helping us do dangerous and mundane things all over the place. That if we wanted to drive we can drive but otherwise you know take a nap or enjoy your car like it's a home theater of yours, you know read from work to home and at that point you're hoping that you live far away and so you could be in a car for longer. And you look back and you realize that there's this company almost at the epicenter of all of that and happens to be the company that you grew up playing games with.

I hope for that to be what the next generation learn. Thank you so much for your time. I enjoyed it, thank you! I'm glad!