Chips are the future of AI. They're also incredibly vulnerable. With Chris Miller

Aza Raskin: Hey, everyone. It's Aza. In 2017, something really miraculous happened, something that has never before happened in the history of humanity, and that is humanity learned how to make machines think, to turn chips or computers, not into just calculators, but into thinkers. From chips to cognition. And this is a big deal, because that means chips and compute is about to become one of the most valuable commodities in the world. Actually, it will likely become the most valuable commodity in the world. And it's a commodity unlike any other commodity because it is intelligence. And so, no matter how much of it you have, well, intelligence can figure out how to use more intelligence. There is no upper bound.

And I cannot overstate how much the US company, Nvidia, dominates the chip race, and how heavily the whole world relies on it. It controls 80% of the entire GPU market, but then all of those chips are made by a Taiwanese company called TSMC. Now, other companies are coming for it, or at least trying. Intel is racing to launch a new AI chip this year. Meta is planning on using their own custom chips, as is Google, as is Amazon. And meanwhile, the US is racing to onshore the manufacturing of its own chips. We don't want to be dependent on other nations, especially ones that are so physically close to China.

The Biden administration just committed up to $8.5 billion under the CHIPS and Science Act so that Intel can build its own brand new chip making centers across the US. And all of this is for this tiny, little object known as the GPU.

And we're going to talk about all of this today with Chris Miller. He is an economic historian. His day job is the assistant professor of international history at the Fletcher School of Law and Diplomacy at Tufts University, and he's the author of the bestselling book, Chip War: The Fight for the World's Most Critical Technology, which I think made it onto the 2023 must read lists of Foreign Affairs Magazine, Bill Gates and Barack Obama.

So, Chris, welcome to the show.

Chris Miller: Well, thank you for having me.

Aza Raskin: Well, let's start by doing some table setting. People may have heard chips, GPUs, microprocessors, semiconductors. We use these terms interchangeably. Walk us through what compute is, when people say compute, and give us a little bit of its history.

Chris Miller: Well, a chip, as it's probably most commonly known as a piece of silicon, in most cases, that has lots of tiny microscopic circuits carved into it.

Speaker 3: Here is a modern 1966 version of integrated circuits, with many hundreds of components on this one circuit. This particular function provides 16 bits of digital memory in this one package. Here is an-

Aza Raskin: In the early days of the industry, when the first chips were invented in the late '50s and early '60s, a chip would often have a handful of components on it. Little switches called transistors that flip circuits on and off. But today, the most advanced chips can have tens of billions of these little transistors, and as a result, they're tens of billions of times more powerful than they used to be. And for any type of computing that happens in the world today, it's represented by these chips, all the ones and zeros undergirding all software, data storage are just little circuits flipping out and off on a piece of silicon.

Somebody who was recently telling me, I don't know if this is right, that the width of the lines etched for when you make these chips are like three nanometers, which is the distance your fingernail grows while I say the sentence. Give us a skit, like, why is it so hard, right? If it's so valuable, are these things hard to make? Why aren't more people making them?

Chris Miller: Well, if you take a chip inside of a new smartphone, for example, open up a phone and look at the processor inside, what you'll find is a chip that has roughly 10 or 20 billion tiny transistors carved into it. And so, to fit in 10 or 20 billion of these devices into a chip that's roughly the size of a fingernail, each one of them has to be tiny, roughly half the size of a Coronavirus, to give you a sense of the scale. And so, manufacturing Coronavirus size devices by the billions is the hardest manufacturing that humans have undertaken. It's more complex than anything else that we make. But in the chip industry, the manufacturing is so complex, so R&D intensive, so costly that there's just a couple of companies that can produce chips at the cutting edge.

Aza Raskin: Got it. So, walk me through that ecosystem. Who can produce chips? What do they depend on? How fragile is it? I just want to understand a little bit of the landscape for everyone.

Chris Miller: Talking about the chip industry, you've got to divide into different types of chips. And so, chips do lots of different things. You take your phone, for example, there's a chip that connects to the Bluetooth, a chip that connects to the Wi-Fi, a chip that manages the camera. But the most important chips are processor chips, chips that run operating systems on your phone or that train AI systems and data centers. And if you look at the ecosystem of companies that produce advanced processor chips, there's really just a couple in the world. In terms of manufacturing, almost all of the most cutting-edge processor chips are produced by one company, Taiwan's TSMC, the Taiwan Semiconductor Manufacturing Company, which manufactures almost all GPUs and a huge share of the advanced processors that go in computers and then smartphones.

But TSMC actually doesn't design any of the chips. They manufacture chips for other companies. And so, whether it's Nvidia or Apple or Qualcomm or AMD, the world's largest chip design companies generally rely on TSMC to manufacture their chips.

Aza Raskin: So, what you're saying is that, many people can come up with the designs for chips, but it is TSMC that ends up actually making the chips. So, they're the final bottleneck. That bottleneck exists in, as you said, Taiwan, which is very interesting from the geopolitics, I think, everyone is familiar with... During COVID, suddenly, cars became hard to get, and cellphones became more expensive because chips suddenly weren't able to get to us and became aware of how, in some sense, fragile or dependent we are. How did that come to be? That seems very surprising that the most important commodity is essentially controlled by just one company.

Chris Miller: Well, in the chip industry, the economics have been defined by economies of scale. The more chips you produce, the lower your unit costs fall, and even more importantly to that, the more rapidly you improve technologically. Because for every chip you produce, you gather data about the production process, you tweak your manufacturing, as a result, and you get better chips in the end. And so, TSMC is the world's biggest chip maker, and as a result, it's also the world's most capable, most technologically advanced chip maker when it comes to producing these processor chips, because it's learning more, simply because it produces more.

And so, that has led it to acquire a market position that is extraordinarily hard to displace. And so, other companies that are trying to compete, like Samsung or like Intel, right now, have much smaller scale than TSMC. And so, they struggle to get the technological advancements that TSMC can gather from their vast production, and they struggle to get the low costs.

Aza Raskin: I would assume that every major country, China and the US, would be trying as hard as they can to undo that bottleneck. Could you walk through what's going on there? US, I think, just invested how many billions of dollars into Intel, to try to make this happen? And again, if there's that much money flowing in, why is it so hard?

Chris Miller: Well, you're right that many countries have looked at their reliance on chips made in Taiwan as a potential vulnerability. From the perspective of Beijing, China spends as much money each year importing chips as it spends importing oil. And Taiwan is one of the largest sources of those chips. And so, the Chinese government, for the last decade, has been in the midst of a process of trying to reduce their reliance on imported chips. But the challenge they face is that, they're meaningfully behind the cutting edge.

And the US finds itself in a somewhat similar position. Key US companies, like Apple and Nvidia, rely exclusively on TSMC to produce some of their most important chips. And the US has been, via the CHIPS and Science Act, which was passed two years ago, trying to incentivize more chip manufacturing in the United States, and it's giving out grants to a number of companies, including TSMC, to open some new facilities in the US. But the reality is that, all of the impact is going to take a lot of time, and it's only going to be a relatively small impact in comparison to the scale that TSMC currently has.

Aza Raskin: I sort of want to get a little into the history then, because it seems like, if I was Taiwan, it'd be no accident to want to have China be dependent on me for something so core to their economy. So, how did TSMC come to be? How did it work in Taiwan's strategic plans?

Chris Miller: Well, Taiwan has been a part of the electronics supply chain for now, well over half a century. But for a long time, it was in the bottom rungs of value add. It was simpler assembly, for example, that was happening in Taiwan, and the more high-tech parts were happening in California or in Japan. But around 30 years ago, the Taiwanese government decided to make a bet on an entrepreneur named Morris Chang, who had spent his career at Texas Instruments in the United States, but was passed over for the CEO job and was looking for something else to do. And he was approached by the Taiwanese government and asked if he wanted to start up a company in Taiwan to manufacture chips. And he had an idea that would transform the industry, and it was to split the design and manufacturing of chips into two different parts.

Before that time, almost all chips were both designed and manufactured by the same company, but he wanted to only do the manufacturing part, letting him manufacture larger and larger volumes, driving down the cost, and also driving up his ability to learn. And that business model innovation explains why TSMC and Taiwan are today at the center of the chip industry.

Aza Raskin: And then, how does that fit into Taiwan's strategic plan?

Chris Miller: Well, you're right to ask, because although it was a business model innovation that made TSMC so important, TSMC is now critical for Taiwan as a whole. In an economic sense, TSMC is by far the largest source of Taiwan's exports. And in a strategic sense, TSMC makes sure that Taiwan is at the center of discussions of technology. And so, today, it's not just China, it's also United States that's hugely dependent on Taiwan, which the Taiwanese hope will give both of those countries an incentive to keep the geopolitical relationship somewhat stable.

Aza Raskin: Chris, let's get into the phenomena that describes the incredible progress that's been happening in chip manufacturing. It's called, Moore's Law. Tell us about it.

Chris Miller: Well, it was named after Gordon Moore, who's one of the two co-founders of Intel. In 1965, he hypothesized that the number of transistors per chip would double every year or two over the subsequent decade, all the way to 1975. And that proved true, and it's proven true with some changes in the rate of growth, but the exponential growth has persisted all the way up to the present. And so, we get chips that are roughly twice as good every two years, and that's meant that the capability of chips has far surpassed the rate of growth in, basically, any other product in all of human history. Nothing comes close to exponential growth sustained over many decades, and it's made it possible to, in an economically viable way, produce larger and larger volumes of computational power, beyond the imaginations of anyone who was producing chips in the '60s or '70s.

And if you look at large AI systems today, they're only viable, only possible to think about because we have access to better and better chips with more and more transistors.

Aza Raskin: And so, do you see any sort of end in sight for this race? Or what are the limits to the race?

Chris Miller: Well, Gordon Moore himself predicted that Moore's Law is going to end at some point. Exponential growth can't go on forever. And so, the question is going to be, when? But we haven't gotten to that point, yet. And indeed, we've actually had repeated predictions from leading experts that Moore's Law is about to end, all the way back to the 1980s, and it's been repeatedly wrong. He was initially focused simply on shrinking transistors and putting more of them on chips. And that's still happening. But in addition to that, we can also design chips differently.

So, what differentiates Nvidia from AMD? Well, it's the way their chips are designed. And you can get more performance out of certain designs than others. So, that's an additional tactic for getting more compute per chip. In addition, you can package chips together differently to have faster data interconnect, for example, between your logic and your memory. That's another way of getting performance. And so, all that's to say, there's a lot of different techniques for getting improved performance, so we're not just relying solely on making transistors smaller and smaller and smaller.

The second reason that Moore's Law has persisted is because there's a huge economic incentive to make it work. Nvidia is the latest example of this. By designing chips better, Nvidia turned a type of chip that was initially used for computer graphics into something that's central to AI. And so, doing made itself a company valued at over $2 trillion. Well, that will incentivize the next company to find another way to improve computing power, and that will keep the Moore's Law dynamic alive. So, I would be someone who would bet on Moore's Law continuing for some time, just because the incentive to find a way to keep it working is so large.

Aza Raskin: Imagine, sometime in the future, someone trains a model that just has incontrovertibly dangerous capabilities, whether it can understand how to make a virus or other biological weapon in novel ways or whether it can copy its code and start self-replicating or... Choose whatever doomsday scenario, you feel like, the red lights start blinking, it has a dangerous capability. When do we need to have put in place the safeties? How long will it have taken, in the past, for us to then make it safe so that, in a sense, you can control anyone else starting to have access to these incredibly dangerous capabilities?

Chris Miller: I think it's almost certainly the case that the answer to that question is not a chip-based question. In other words, the chip will do what you tell it to do.

Most chips are fairly general purpose in nature. And so, what that means is that, if you want to put guard rails up around a system, the chip is not the level at which you're going to do it.

I think that the economics are such that there are strong incentives to produce general purpose processors, which means that we're not going to have, I don't think, in a widespread fashion, chip-based guardrails, it'll be system-based guardrails that you train on a chip.

The second question though is, for deploying a system, how much computing power does it deploy? We know training is super computationally intensive, but inference will be as well, depending on what scale it's happening at. And so, you're going to need a lot of chips, you already do, and you'll need more for inference. And so, that's the second part of the equation is, depending on what type of system you're talking about, what's the computational lead for inference and how easy is that to access?

Aza Raskin: Just jumping in here to orient. For AI, there are generally two portions of training and using an AI. When you train an AI... Well, that's called training. That uses a lot of computation, a lot of power. And then, when you use that AI, when you ask it a question, when you ask it to do something, that's called inference, and it happens much faster, and it generally takes less power but you use it a lot more.

I'm curious, where do you see your greatest concerns? And you've just said, you don't think it's going to be on chip, but in systems. I'm curious to expand that as well.

Chris Miller: If you look historically, whenever a new technology emerges, that is fairly general purpose in its nature. And so, it can be used for lots of different ways. We always don't know the ways it will be used. It's very difficult to predict. And so, there's extraordinary uncertainty about, how do you set up the types of guardrails that you want? Because it's difficult to set up guardrails for something that you can't predict how it'll be used. You set up guardrails for systems that you know how they'll be used. On a road, for example, you know if the road's going to turn left, and so you put guardrails so a car doesn't miss the turn and fall off the cliff.

And so, where we are with AI right now is that, it's extraordinarily difficult, basically impossible to predict all of the use cases that will be envisioned. And so, I think that's why the conversation about guardrails is so conceptually tricky. You just don't know where exactly to be putting them up.

I think the second challenge is that, like all other general purpose technologies that we've seen historically, AI will be used for a lot of good things and a lot of bad things. And so, it will be shaped not just by the technological attributes, but also by the social and political context in which it's being deployed. And so, I think if you're going to think about ways in which the technology becomes problematic, you've got to both think about the technology itself, the process of putting up guardrails around it, and then the social and political incentives about who's actually designing the guardrails. And so, all that's to say, I think it's not just a question of technology, it's also a question of society.

Aza Raskin: Totally. It's something we often say, and this is a quote from Charlie Munger, is that, "If you show me the incentives, I'll show you the outcome." And I'm sort of like ground zero for understanding this because I'm the hapless human being that invented Infinite Scroll. And when I invented it, I invented it because I'm like, "Oh, we reached the end of a blog. If you want to see more blog posts, you reach the end of some search results, you want to see more search results just to load more in." But I was blind to the way that it would be the incentives of the race to the bottom of the brain stem for attention that would turn to use Infinite Scroll as a weapon to keep people versus helping them.

And so, I think the key thing with AI is not to ask, is AI good or bad? It's to ask, is the incentive that will pick up and wield AI, good or bad? I think the thing that people often forget about AI is that, there is no way of separating the promise and the peril. It's the exact same tech that lets you make incredible AI art and decode brain scans into what somebody is seeing. That tech is the exact same as it is to make fake child pornography. You can't disentangle those. The ones that make great biology and science tutors are the ones that make tutors that help you create bioweapons. You cannot separate the promise and the peril, which makes the governance question as you're pointing out incredibly difficult.

Chris Miller: Yeah. No, I think that's right. And that's why we can't, I don't think, count on technological solutions to what are essentially social and political challenges.

Aza Raskin: One of the frames that we'll often use, and we've used it on this podcast, is, what oil is to physical labor, AI is to cognitive labor. That is every barrel of oil is worth, roughly, 25,000 hours of physical human labor time. And after we figured out how to harness oil, you could take people out of the fields and replace them with mechanized workers, essentially tractors. Same thing is going to happen here now with cognitive labor. What is cognitive labor? It's like, when you sit in front of your computer and have to write an email... Writing email, choosing what words is cognitive labor. Deciding, as a scientist, what experiments you're going to run and how to write them up as papers, that's cognitive labor.

And so, it sets up very much another race, like an industrial race between companies as well as countries. And I think that's why we see the US doing their export controls on chips. And so, I'd love for you to... Feel free to disagree or add more nuance to that analogy, and then talk about its geopolitical implications.

Chris Miller: No, I think that analogy is absolutely right, and it's a race with technological and economic, but also defense and intelligence ramifications, too. And so, if you talk to people who work in defense ministries or intelligence agencies, they're also asking themselves the question, how will AI change my work? And the answer is, it's going to be pretty important and pretty impactful. And so, there's a race between militaries and intelligence agencies around this exact topic, and that explains why the US has been trying, over the past couple of years, to limit China's access to high-end compute. And so, starting in 2022, the US imposed restrictions on the highest range of GPUs.

Aza Raskin: And just a reminder for people who might be lost, GPUs are built on chip. So, they're a specific kind of chip surrounded by a certain kind of architecture, originally made for gaming.

Chris Miller: Then, in 2023, brought the restrictions a level further down to sort of the second-highest range of GPUs, making it illegal to transfer them into China, with the aim of making the cost of compute in China higher, adding inefficiency to the AI training and deployment process, and the US hopes keeping a US edge in AI over China.

Aza Raskin: I was just going to ask, do you know how effective that's been? We've heard, through some of our friends that work with Chinese researchers, they feel like they're at least a year behind and they're frustrated by the inability to get chips. But we've heard from other people that you can get them on the black market, and the chips get sold to some other country and then imported into China. So, do we know how effective it's being?

Chris Miller: It's a tricky question to answer because any sort of way around the controls is naturally not going to be recorded and publicized. It's also tricky because I hear about shortages of GPUs in China, but I hear about the same thing in Silicon Valley. And so, it's difficult to know for sure that... There certainly is smuggling happening. There's been well-documented instances of that. My sense is that, on balance, it is creating some challenges for Chinese firms to train at the level they like to train and driving up the cost, adding some efficiencies.

On its own, it's not going to stop China from training any individual system, because you can just train a system with a less efficient data center. It takes a bit longer, but you can do it. I think where it's going to have a bigger impact is on the amount of training, especially not by government actors, but by business actors who are going to be more responsive to the price of training relative to the government.

So, it's going to be a mixed and complex picture, but I think it is having some impact right now.

Aza Raskin: You studied the role of technology and the downfall of the Soviet Union, and so a question I have is, is it more important for the US to stop China from developing powerful AI for its military without necessarily developing that power ourselves? Or do we have to do both? That is to say, if our goal is to beat China, and we hear that again and again in the halls of Congress, "We must beat China", let alone the fact that we beat China to social media, which means we actually lost when it came to polarization and mental health, and a whole bunch of other things we really care about to strengthen our society. What's required to "win" with China is just slowing them down enough?

Chris Miller: Well, I think that the reason this is a tricky question is because nobody's really sure what the most useful application of AI will be in military contexts. Right now, militaries are experimenting, they're deploying AI and intelligence systems, for example, to gather pictures and use computer vision to identify what's dangerous and what's not, or to make sense of signals intelligence data.

It's being deployed to make systems more semi-autonomous, but no one's really sure exactly how it will be the most impactful. And what that means is that, both China and the United States are betting not just on the deployment of individual AI systems, they're also hoping that their AI ecosystem will be more developed and give them a broader level of capabilities to draw on in the future. And so, that's why there's both sides of this race trying to slow China down, but also trying to race ahead, are the strategies being deployed right now in Washington.

Aza Raskin: One of the goals for many of these AI companies is to automate science itself. That is, essentially, have the AI to discover new types of materials which, if you can discover new kinds of batteries, you can discover new kinds of bombs. If you can discover new kinds of cancer drugs, you can discover new chemical and biological weapons. And so, I'd imagine, that's another thing deeply driving this race is, whoever can get to a increased rate of science, dominates.

Chris Miller: Yeah, I think that's right. And the perception among policymakers is that, there will be meaningful productivity improvements to the economy and also to research processes in particular, which mean that it's not just about the question of, what will the impact of AI be on our economy? But also, if you're deploying AI more rapidly than your competitor, your economy grows more rapidly, you're in a better strategic position, as that is, again, another incentive to run in this race that you alluded to.

Aza Raskin: Yeah. I'm always very hesitant to reify talking about any kind of war, especially war between the US and China, because we have nukes now, and so the stakes are much, much, much higher. But it reminds me of that moment when the US, before World War II, cut off our export of oil to Japan, which I've heard some historians cite as a galvanizing reason for why Japan got into World War II.

And I'm curious, does China view the export controls as some kind of beginning of an act of war? How do they view that? What are the dangers that you see here?

Chris Miller: Well, I think you're right that there is a danger of pushing China in the direction of thinking it's got no choice but to turn to war. The context in which China make that type of decision is, if it believed it, it was in a race but it could not win, and therefore it was better off taking a risk now rather than taking a risk in a couple of years time when it was in a worse position.

I think there's some key differences, though, between today and the episode in the 1940 to '41 period that you were discussing, pre-World War II. One is that, in Japan 1941, had they not gone to war and had the oil export bans remained in force, Japan would've run out of oil and their economy would've frozen in 1941 or early 1942. For China, right now, if it can't access cutting edge GPUs, there might be substantial long run costs, but the short run cost is pretty limited. China's economy is doing badly, but not because they can't get GPUs for its own internal reasons. And so, there's not the sense of crisis that the oil export bans created to Japan in 1940 to '41.

The second dynamic is that, Japan knew it couldn't synthetically produce oil domestically. Chinese leaders think they're going to catch up in terms of producing GPUs domestically. And I'm skeptical, but I think they've got a shot, and that, I think, is to them a better bet to take than rolling the dice on a very risky war that would, as you say, be disastrous for everyone.

Aza Raskin: How do you think this is going to change as the capabilities of the new models get to be known? So, the estimates that we're getting from across the different labs is that we're around nine to 18 months away from the AI models having the capability of programming at roughly the human level. That is, you don't have to hire a Google engineer anymore, you can just spend the money, use the compute to generate, essentially, a synthetic programmer.

And of course, what happens there is, this is the first time, really, that human beings and money become fungible, because before, you would have to hire people, find them, train them in your culture. Now, you just pour in money and you get out digital programmers, they increase your capacity, they make you more money, which means you make more programmers. And it seems like, as soon as that happens, which is very soon, the countries that have access to asymmetric amounts of compute, start to asymmetrically take off. And that seems, at that moment, China might feel themselves falling further and further behind.

Chris Miller: It's certainly possible. I think if you talk to policymakers in western countries, and I think the same is true in China, although it's a bit more opaque. I think in, basically, all policy contexts, political leaders really struggle to see around technological corners. They're not technologists. And anyway, predicting the future is very hard, especially around points of technological discontinuity.

And so, the assumption that they naturally make, is that things will look pretty similar in two years and in four years with marginal change. And in most facets of life, that's a good assumption to make. And at points of technological discontinuity, that's a bad assumption to make. But I think that is the reigning assumption among almost all policymakers, not because they're thinking carefully about the problem, just because that's the natural assumption to fall into.

Aza Raskin: Yeah. And this is such an important point because even the people who work on these AI systems every day, are making systematic errors in their predictions for how fast things come. And the systematic error that they make is that they are constantly getting it wrong in the direction of things happening much faster than they think.

And we're seeing that across the board. Every benchmark that the AI industry sets, we are hitting those benchmarks much, much faster. And that's because even the latest Nvidia chips, the H100s, we are starting to get to the place where AI is used to make AI better. So, I believe that AI was actually used by Nvidia to make their H100 chips more efficient, which makes their AI more powerful, which lets them make even more efficient chips.

So, we're sort of starting to enter this double exponential. What do you see as the ramifications of policy makers, sort of rubber banding back into more of a linear mindset rather than this exponential or even double exponential mindset?

Chris Miller: It's tricky to think through the ramifications of an AI acceleration on the chip industry, because as you say, on the one hand, it is the case that both chip design companies, the companies that make the software tools that are used to design chips, and also chip manufacturers, they're all using AI to try to improve the processes. On the other hand, the better AI gets, especially if it happens rapidly, relative to expectations, the more chance there is that you have disruptions to existing business models, and potentially even a challenge to the market positions of leading firms.

For example, if you believe that AI will be at human levels of programming in X number of years, it will probably also be at human levels of chip design. And so, what's the ramification of that? Well, that's very difficult to say with any confidence, but I think it makes projections like this quite difficult.

Aza Raskin: So, refocusing for a second on Taiwan, because I think if you had asked people a couple of years ago, they would've said, "If China were to attack Taiwan, in the end, the US would win." But now, China's name has gotten bigger, China has the home court advantage, especially with supply chains. Given Taiwan's central role in producing chips, combined with China's desire to get AI supremacy, does that make war more or less likely?

Chris Miller: Today, China relies on chips made in Taiwan just as much as anyone else. And so, if there were to be a war, Taiwan's chip making facilities be knocked off-line on, basically, day one, and so China would lose access to chips, not only for AI but also for all types of manufactured goods.

Aza Raskin: And when you say knocked off-line, what do you mean? Because if somebody would blow it up intentionally... What do you mean by that?

Chris Miller: Well, you don't really even need to blow it up intentionally. That could happen, but if you just constricted energy flows into Taiwan, Taiwan imports much of its energy via liquefied natural gas. That alone would be enough to shut down Taiwan's chip-making facilities. They also import lots of chemicals from Japan and elsewhere. So, any type of blockade scenario would result in the facility shutting down. And so, they're, as a result, quite vulnerable to any sort of geopolitical escalation.

Today, everyone would suffer in, perhaps, varying degrees, but everyone would suffer if we lost access to chips made in Taiwan. But China is trying to become much more self-sufficient, and it's going to make progress towards that goal over the coming years. And so, I think by 2030 or so, China will have a lot of the chip-making capabilities that it needs domestically. The question is, will it have the high-end capabilities that are needed for producing AI chips, for example, at scale? That's uncertain. It depends on the rate of technological progress in China. But if it does, then it would face less costs from knocking TSMC off-line.

Aza Raskin: It feels like we're essentially heading into another kind of Cold War here, like a compute Cold War. And I'm curious, what are the lessons, if any, that we can draw from our last Cold War with the Soviet Union for how that might play out this time?

Chris Miller: Well, the Soviets realized, to some extent, they were in a compute Cold War as well as a real Cold War, but they didn't really have a good strategy for competing in it. They were fixated on copying U.S. technology rather than developing their own. They never scaled up production domestically, and as a result, their compute resources were hopelessly inefficient and behind the technological curve.

China's trying to avoid that mistake by investing in their own chip industry, by trying to scale up using their vast domestic market. And so, they're still behind, but they're much closer to the cutting edge than the Soviets ever were. So, that's why the US is fixated on this issue of compute in a way that it really hadn't been for several decades, because it's got to compete now, because the competition is really so close.

Aza Raskin: As we're just talking about, as AI gains ability to code or to replace human cognitive labor, those countries that have access to compute, obviously outcompete the ones that don't, and starts to increase wealth inequality at the country level and the person level. I'm just curious to hear you expand on that.

Chris Miller: Well, I think the dynamics are actually complex, and perhaps more contradictory. If you ask yourself, why are wealthy countries today, wealthy? It's because their cognitive labor is highly valued. And so, insofar, as machines are capable of substituting for some of that, actually you might have some pretty contradictory facts.

Aza Raskin: You mean that people that currently have high paying jobs, suddenly will be out of those jobs, and so the internal dynamics of that country might get very challenging? Is that what you're saying?

Chris Miller: Well, that's right. And if a country has a high level of income due to cognitive labor, and suddenly cognitive labor is mechanizeable in a large scale way, this undermines their business model.

And so, I don't think it's necessarily obvious that at a country level AI, if it's successful at replacing large numbers of cognitive tasks, that it necessarily benefits the countries that are on top right now. It could benefit certain groups in those countries and others, but it seems like it's a much more complex story than just the sort of winner-takes-all at the country level.

I think the other aspect is that, if you think of the two main inputs being compute and power, we've talked a lot about compute, who produces the compute, also who produces the power becomes pretty important, where power efficiency dynamics trend. And it's been interesting. Countries in the Persian Gulf, for example, like Saudi Arabia or the UAE, try to play a bigger and bigger role in data center construction, for example, arguing essentially, we've got the power. You can put your compute anywhere in the world, but our power resources are what's actually limited. And that's something you, I don't think, would've expected if you were to ask about, what's the economic implications of more and more capable AI? But if, in fact, power is one of the limiting factors, that could be something that ends up being pretty important.

And so, there's a lot of focus right now, for example, in cooling data centers more effectively. The more you cool them, the less power they draw. And also, in designing chips in different ways, so that they use less energy per unit to compute. And so, I don't have a high confidence view as to what the slope of that line will be, but it's a pretty important question.

Aza Raskin: What keeps you up at night? When you send your mind to all the places that things could go poorly, what is at the top of the list for you?

Chris Miller: Well, I do think that the chip shortage that we experienced during the pandemic was just a tiny fraction of how bad things could be, how disruptive things could be if we were to lose access to Taiwan. It's not just the high-end data centers that are used for AI, it's medical devices, it's cars and it's household goods. And so, the entire world manufacturing sector today is tied directly to chips that are largely, not exclusively, but largely made in Taiwan. And so, the world economy sits on top of this, I think, pretty fragile foundation of silicon, which, for a very long time, we've taken for granted.

Aza Raskin: I don't know exactly how to ask this question, but maybe I'll just start by asking you, where are your worries about the AI capabilities themselves? Is that something that worries you of what they might do? What proliferation of intelligence into society without bounds sort of like amoral intelligence might do? Or is that not where you placed the biggest risk?

Chris Miller: Well, I guess, I think if you look at the history of computing, you've got to conclude that more and more capable systems will inevitably follow. The economic incentive to create them is so large. The competitive dynamics, the race you alluded to is so fierce that they're going to be created. And so, the question is, how do you produce the right context in which they would be used for better and not worse outcomes? And that gets back to the question of, what are the social and political incentives? What are the economic incentives around them? And what are the types of guardrails that you want?

But I'm not someone who believes that technological improvements are going to be stopped, and that's just not going to happen given the incentive to keep building. And so, the question is, well, what's the context in which you want this building to happen?

Aza Raskin: Right. It's a little different, but the feeling is, instead of saying, "We must prevent climate change, we must prevent climate change", it's like realizing that we're in mitigation or adaptation. Like, "Climate change is going to be here" or "We're going to do something about it." So, what I'm hearing you say is that, the history of compute says, the physics and the incentives of it are that we're going to get faster chips, we're going to get them at cheaper costs, we're going to discover new physics that accelerates this whole thing, not just like 10 X, but 1,000X or more. And then, it's about figuring out how to mitigate the consequences, rather just preventing it.

Chris Miller: Yeah. And I think promising a 1,000X improvements in most industries sounds like a pretty big promise. But in the chip industry, a thousand X, we've done several times before.

Aza Raskin: Well, in fact, since the beginning of chips to now, what is the total multiple of increase in speed that we've seen?

Chris Miller: There are different ways you can measure, but if you take... The first commercially available chip had four transistors, and so there's about 10 billion times the number of transistors on a new Nvidia GPU.

Aza Raskin: Is there any other industry that has had similar rates of improvement?

Chris Miller: No, nothing else comes remotely close. And so, I think that speaks to the challenges of trying to do on-chip governance. There are efforts underway to explore ways to make this robust, but I think like any sort of restriction you put on a piece of hardware, it's only as good as the incentive to break it is small. And so, that, I think, is one.

I think the second thing is that, it's easier to envision straightforward limitations, but harder to envision more sophisticated ones. And so, you can't go above a certain speed or you can't go in a certain location, that's relatively straightforward. But saying you can't train something that's dangerous is much, much more difficult to envision how you'd even begin to go about that type of process.

Aza Raskin: Chris, I have to ask you now a final question. And this one comes from Sasha, our executive producer. And her question, I think, is one that many parents have, which is, given the direction this is all going, what is the future prospect for our kids? What should they study now to have them have any kind of job in the future?

Chris Miller: Oh, I don't know. That's a hard question. Well, I guess, I think the history of most big technological shifts suggests that, what you want above all is the ability to harness those for economic purposes, and you don't know what the right way to harness those are going to be.

But in terms of skills, I think the skills that are going to be the most valuable are the skills that let you manipulate technology to do useful things with it. And so, that probably doesn't mean traditional stem, just like I think my generation practiced a lot less long division than my parents did, because we knew we could rely on calculators. Maybe the trajectory of change is going to be more rapid, but I think that the basic dynamic of trying to find ways to harness technology, the skill sets you need to do that, will be the source of economic value in the future.

Aza Raskin: Awesome. Chris, thank you so much for coming on *Your Undivided Attention*. Understanding chips and GPUs and geopolitics. It's a lot to hold in your head, but it's also critical to understand where we're going. So, thank you very much for elucidating it all for us.

Chris Miller: Well, thank you for having me.

Aza Raskin: All right. Just one more thing, one more thought. Chris doesn't believe that on-chip governance, the ability to control how much compute is being used for what, is going to be practical. Now, I don't know if that's true or not. I've heard other opinions. But if that is true, it means the cavalry isn't coming. We're not going to get a technical solution to choosing how, in what ways, compute that is cognition, that is intelligence will be used by humanity.

What that means is that, over all governance, trustworthy governance, is where the solution is going to have to come from. And I think something we could probably all agree on is that, if hundreds of billions to trillions of dollars is going into creating AI capabilities, then at least one to 10% of that should be going into figuring out what is the form of trustworthy governance that steers the entire thing.

Your Undivided Attention is produced by the Center for Humane Technology, a nonprofit working to catalyze a humane future. Our senior producer is Julia Scott. Kirsten McMurray is our associate producer. Sasha Fegan is our executive producer. Mixing on this episode by Jeff Sudakin. Original music and sound design by Ryan and Hays Holladay. And a special thanks to the whole Center for Humane Technology team for making this podcast possible. You can find show notes, transcripts, and much more at humanetech.com.

If you liked the podcast, we'd be grateful if you could rate it on Apple Podcasts, because it helps other people find the show. And if you made it all the way here, let me give one more thank you to you for giving us your undivided attention.