Tech Summit on Artificial Intelligence: A Quote Book

Hardware and Infrastructure Edition: Semiconductor Chips & Cloud Services

March 2024

US Federal Trade Commission
Office of Technology

Table of Contents

Overview

On January 25, 2024, the FTC held a Tech Summit on Artificial Intelligence. The event page, with the full 4.5 hour recording of the event, is available [here].

In the first panel, we hosted the following panelists:

- Tania Van den Brande, Director of Economics, Ofcom UK
- Dave Rauchwerk, Technologist and Entrepreneur
- Ganesh Sitaraman, Director of the Vanderbilt Policy Accelerator
- Corey Quinn, Chief Cloud Economist, The Duckbill Group

Panel Summary: The panelists discussed ways that dominant firms may have control over key infrastructure inputs such as cloud computing and access to hardware such as GPUs, and this may be exacerbated by obstacles to migrating between offerings. This may in turn allow them to charge excessive prices or impose coercive terms, and as a result, they may be able to exercise market power in ways that favor their own incumbency or impact competition. In addition, the panelists discussed ways in which the structure of these markets may make it challenging for new players to compete even where their products and services may be better than the incumbents.

Why a Quote Book? The voices of everyday Americans can sometimes be lost in discussions involving dense technical, policy, or legal language. While the benefits or risks of new technologies are being debated by policymakers, these individuals experience the effects of innovation in real-time.

The FTC recognizes that this is not a representative sample of the entire population, and we strive to continue to listen and engage with a variety of perspectives. The goal of the quote book solely aims to reflect and compile quotes from the participants aggregated into common themes. This summary aims to be a resource, to quickly see various perspectives on topics.

Ideas to enable competition and innovation

"It's really interesting that the FTC has convened this because we're in this moment where, okay, what is the point at which the market leader who gets there by fair means over 30 years becomes anti-competitive? And what needs to change so that we can have more companies building more chips?" - Dave Rauchwerk

Addressing vertical integration through structural separation

- "So one thing I think that's important is we want to see innovation. And I think a critical part of that when you think about the lower layers of the stack is the possibility that we will have innovation be foreclosed by vertical integration and dominance at these lower layers. And so I think addressing that is a critical issue." - Ganesh Sitaraman
- "But the first is a structural separation. And this is basically an idea that you restrict the lines of business operations within an entity so that they can only operate one or a set number of lines of business as opposed to being vertically integrated across many different lines of business or a conglomerate that applies across many different lines of business. That is a very clean and administrable way to prevent things like self-preferencing and other kinds of harms that may emerge from integration." - Ganesh Sitaraman

Regulatory action

- "So in the UK context, currently, our competition authority is looking at these issues and to the extent that it finds concerns, we'll be able to make interventions, and you can imagine that that might focus particularly in part on some of that ability and that incentive for customers to switch." - Tania Van den Brande
- "I think regulators and authorities, not just in the UK but across the world, I think are starting to look into the risks that you've been highlighting during this talk, and I think have several tools available to them to intervene where we can find solutions. And particularly in UK, we're about to start doing regulation of the big tech companies on an ongoing basis where some of the issues, if they do arise, could be dealt with on a more ongoing basis. So really value the discussion today to really help the creative juices in how we might tackle issues as they emerge." - Tania Van den Brande
- "I think there's a real danger from concentration to innovation in these downstream areas based on dependence on the more utility like elements upstream. And I think that's a real concern. And so solutions to addressing that problem in a combination of enforcement and regulatory actions, I think are very critical. And we could talk about that more later or now if that's of interest. But that's a place where I think we should be very worried." - Ganesh Sitaraman

Non-discrimination rules

- “Second are non-discrimination rules. These are rules that say first that an entity that is operating a kind of service has to treat everyone on similar terms and really create a level playing field that again, would apply to self preferencing, but it also applied to other kinds of preferences that could be both related to price, it could also be related to terms and conditions or different other kinds of types of orders or applications. So I think having a sense of non-discrimination rules is another important way to ensure that there's confidence for users of a service that they are not going to be effectively price gouged or appropriated out of all the potential benefits that they might have for their innovative idea to make profits and money by the entity that has a bottleneck over an essential service.” - Ganesh Sitaraman

Interoperability

- “...you need to have a possibility for customers to move around easily and to benefit from those innovations that sort of fits what they need.” - Tania Van den Brande
- “[I]t's got to be easier for customers to move around and choose a cloud provider that has a set of AI solutions that they're most interested in, and not necessarily just the ones from the cloud provider themselves, but from third parties that might build those apps on top of those clouds.” - Tania Van den Brande
- “[W]hat kinds of rules there are around interoperability. Again, something we haven't talked as much about, but the ability to actually be able to switch between different providers. So I think those are some of the things that are kind of standard legal tools both in the remedy context in antitrust cases or in the regulatory context that we've seen imposed in many different sectors where there are similar concerns about monopoly or oligopolies being dominant and the harms that can come from them.” - Ganesh Sitaraman

More competition and more options

- “So it's this thing where if you want to have more chip companies you need more cloud companies, and we have too few cloud companies and now they're making their own chips. So we're sort of stuck. And it's this thing where real innovation is possible, real innovation exists in the market from the very fundamental base layers of the stack.” - Dave Rauchwerk
- “There is the possibility for a future where there are more clouds, where there are more chip companies, where the open source community has more options than just Nvidia, where an AMD can say, you know what? It's actually worth prioritizing even more investing in the ROCm and making sure that there is portability. Those things will happen, but we have to take action, we have to do something about it.” - Dave Rauchwerk

Increased transparency

- “I think that in the short term increased transparency because right now it's who can genuflect the hardest to the feudal lord that they are sworn to. That's not a viable system for modern governance around distributing an asset that is right now incredibly rooted in scarcity. I want to see a better tomorrow, not so much a better 10 years from now. How do we start making small steps today rather than hoping for an ultimate solution decades from now when it's even harder to change then?” - Corey Quinn
- “A third and related point to non-discrimination rules is transparency of some of these terms and conditions so that we really know, and I think this is something that Corey has referenced a couple of times, that in some cases we may not even know what is going on in some of these areas.” - Ganesh Sitaraman

Semiconductor Chips

The Founder Perspective: Startup challenges in the AI space

- “As a startup founder, you are basically competing with the most valuable companies in the world. And it goes beyond the success of Nvidia, which we can talk about. And it includes fundamentally all of the major hyperscalers, all of the major platform companies. What we've seen over the last five years, through a series of acquisitions, is platform companies, hyperscalers, starting to make their own chips.” - Dave Rauchwerk
- “So this is Amazon, this is Microsoft, this is Meta. Even Tesla is making their own chips. And what this does is, it further makes it difficult for new entrants to come into the market. And on top of that, when you look at, as a startup founder, you look at the access to capital, you have to go into the room and explain to investors, "Okay, so we're going to compete with the largest companies in the world, and we're not sure if they're going to ever buy our chips because they're already making their own chips."” - Dave Rauchwerk
- “If we all want to continue to lead in AI, which we do, we need to have more companies than just the few, and just the hyperscalers, producing these chips. So in order for there to be a market for these chips, there has to be investment. And the problem now is, I looked at the data recently, and there's about 5,000 venture capitalists who have made investments or actively investing in AI startups, but there's only 300 that are investing in semiconductor startups.” - Dave Rauchwerk
- “And so when we look at this, it's this situation where the dynamics of the market basically make it a non-starter not just for entrepreneurs, but also for the investors themselves. Now, this is not to discount the success and the emergence of some really great AI-focused semiconductor companies, but they have primarily had to compete with hyperscale companies that have enormous volumes, and themselves can finance the development of their own chips.” - Dave Rauchwerk

- “So, as a former semiconductor founder, now has never been a better time to be in the semiconductor business. We're going to have more fabrication capacity online in the next five years within the country than we've ever had. And there's enormous demand. And yet, the dynamics of the market make it extremely challenging to get off the ground.” - Dave Rauchwerk

Security Risks: Single points of failure

- “Oh, dear Lord. I think the war has already been lost. We've passed a tipping point where you cannot avoid the three main hyperscalers out there, full stop. In fact, if there were to be a law or technical issue passed tomorrow where AWS could onboard no new cloud customers, they would continue to grow revenue for at least several quarters just based upon organic growth.” - Corey Quinn
- “Too big to fail has passed a tipping point long ago, and the centralization risk is massive. Once upon a time when we all ran our own data centers things went down a lot more, but the failures weren't correlated. It wasn't effectively every business having a problem.” - Corey Quinn
- “Even today, if you decide that you want to build an e-commerce store and I'm going to build it on Azure so I don't have to deal with AWS in any way, well, if you're using Stripe to handle your checkouts, they're a full in AWS company. So if AWS has a bad day, no one's buying anything on your shop. Those dependencies wind up happening across the board.” - Corey Quinn
- “In the chips layer, the CHIPS and Science Act is addressing a different kind of resilience challenge, one tied to geographic production of chips and where they're located, obvious national security issues and concerns there. So I think one of the other places we need to think is just how concentration can be important in that direction.” - Ganesh Sitaraman
- “On a couple of other points, I think if you are the federal government, there's also a question of dependence on a single actor in the private sector for a significant amount of your compute power or really for any other resource as a federal government, one of the challenges that you worry about in a democratic republic like our own is who's actually calling the shots. And to the extent that there's significant lobbying, regulatory capture and other dependencies, it may be that government actors over time don't feel like they can take significant enforcement or other actions against companies on which they are dependent.” - Ganesh Sitaraman
- “The government, the US federal government runs a staggering percentage of its compute workloads on the big three hyperscalers. I'm not suggesting that there's unfair influence of stop investigating us or your computers are going to stop working. I don't think anyone is getting to that point. But there is a sense of how much enforcement can really be done when effectively you are critically dependent upon the continued existence and wellbeing of these companies just to go about the daily business of government.” - Corey Quinn
- “This is the sort of too big to prosecute idea that emerged in a number of sectors after the financial crisis in which there were concerns that the too big to fail, banks were so large and so important systemically in the economy that enforcement actions might be

problematic against them for the effects that they would have. I worry also in that case that we may end up in a similar kind of situation if there's a very limited number of actors upon which there's real dependence, particularly by the government, but across the economy as well." - Ganesh Sitaraman

Competition Concerns

Less innovation

- "[W]hat we might see over time then is actually less innovation in the model or application layers, anything dependent on these concentrated layers where there's a kind of bottleneck of players that can incorporate those new ideas directly into their own offerings and then spread them throughout their vertically integrated business lines." - Ganesh Sitaraman
- "But right now, there's not enough cloud providers and there's not enough chip companies. And if we had a world where there were more clouds and more chip companies, there would be more competition." - Dave Rauchwerk

Self-preferencing

- "A second way that this could happen is not involving any sort of copying or taking of others' ideas, but simply self preferencing one's own integrated business lines over others. And when you're a very large company and you have a lot of customers and a lot of users, you have the ability to preference your own integrated offerings. And what that means is that others can't get a real opportunity to get into that market. And that puts real limits on new providers that might be more innovative, have better ideas, have more interesting ideas. And so self preferencing in this area could be another concern." - Ganesh Sitaraman
- "When you look at the way that a hyperscaler can operate, when they're making their own chips, it gives them unparalleled access to surveil. And a sort of form of innovation surveillance where they can see what their customers are doing. They can look into the memory inside of the chip itself, and they can see what is actually running on it. And what this means is, they can figure out what needs to be made before it needs to be made." - Dave Rauchwerk
- "Concentrated power means that entities have the ability to preference their own vertically-integrated business lines to discriminate against users of their service, and of course to increase price and reduce quality of services." - Ganesh Sitaraman

Tying

- "Traditional issues like tying are another potential problem under the antitrust laws. Obviously tying not allowed, but there have been, in our history, many, many, many examples of companies that have tried to do so anyway, and also enforcers who have gone after them for engaging in that kind of behavior." - Ganesh Sitaraman

Chilling effects

- "This is nowhere scarier than with Amazon itself. Take a look at, what industries does Amazon not operate in? The only one I can think of is philanthropy. Employees in

jurisdictions where it's not barred have to sign an 18 month non-compete scope to Amazon, which means that there's no industry they're not in, there's no way to not run afoul of that, which causes a certain chilling effect." - Corey Quinn

Inability to change providers

- "Moving from a data center into a cloud provider is a massive project that's measured with multiple calendar years. Moving from one cloud provider to another is almost that same level of difficulty. Once you're there, you tend not to move. There are a few stories of people fully leaving a cloud provider that they have been all in on. And for good reason, it simply doesn't happen at most. A workload or two will move from one to another or something greenfield will be spun up somewhere else. But once a workload is there, it basically is there to stay." - Corey Quinn
- "Inertia is such a powerful force it's hard to overcome. Oh, a new vendor to get through all of my procurement processes, my security validation, understand how it works, more importantly understand how it breaks because when it breaks, you really wanted someone who's been there before. And it becomes this almost insurmountable series of obstacles to the corporate decision-making process where, okay, let's put this on Amazon too is a straight shot. The big get bigger and the gulf grows wider. It becomes a bimodal distribution whether we want it to or not." - Corey Quinn

Bottlenecks

- "[At] the lower layers of the stack, there are already monopolies and oligopoly, and concentration is already a reality in the lower layers of the stack. So that's the first point." - Ganesh Sitaraman
- "From my part of it, I think that the biggest challenge that we're seeing is that all roads lead to Nvidia. They are today a bottleneck on all of this, followed only slightly by the large cloud providers that are their primary customers. I think that fundamentally, since all these companies love to talk about being utility computing, it's time to start treating them like a utility. We would not stand for an electricity provider or a water provider that suddenly slapped a zero or two on the end of every price that they were charging people. But there's very little stopping companies from doing that today other than not wanting to destroy goodwill and then evoke a sharp regulatory response immediately. It's instead they're taking the boiling the frog approach." - Corey Quinn
- "At the chips layer, there's one company that is predominant in designing the most advanced chips. There's one company that is predominant in manufacturing those chips, and there's literally one company only that produces photolithography equipment, which is an essential input into manufacturing those chips." - Ganesh Sitaraman
- "I think we've heard today that there are significant problems that come from having concentration at these lower layers of the stack, including to innovation and to having a robust ecosystem for startups and for new entrants." - Ganesh Sitaraman
- "Well, even in the open- source community, in the hobbyist world, everything is around Nvidia. The only time you see people building these things on other chips is similar to the old trick of installing NetBSD on a toaster. Just to prove it can be done, but no one's

seriously suggesting you go ahead and run your data center on those things. It is an Nvidia monoculture now, and that's frankly what scares me." - Corey Quinn

Partnerships vs. Acquisitions

- Now, in the semiconductor business, it is the apex of human science and technology collaboration and coordination. It survives on robust and strong partnerships. Nvidia has been a partner with TSMC since 1998, and it has been an enormously productive partnership for both companies. And the challenge here is when the partnership ceases to be a partnership, and it becomes a competition between a company and its own customer. - Dave Rauchwerk

Cloud Computing

Competition Concerns

Egress Fees

- "And first of all, one of those barriers we saw was egress fees. These are the cost that customers pay to move their data out of a cloud. And according to the work we did, it can really get quite expensive for a customer when it's running a multi-cloud architecture, particularly if it's moving data between different clouds in that process. And we also saw scenarios where those egress fees can make it really expensive, the switch, particularly if that switching needs to happen gradually and customers using a multi-cloud architecture during that switching process." - Tania Van den Brande

Switching costs

- "And the first one of those was the observation that cloud in the UK is very concentrated towards AWS and Microsoft. And secondly, we were starting to see some evidence emerging that customers were struggling to switch." - Tania Van den Brande
- "Now, speaking to egress fees as well, there's a misunderstanding in many parts of the industry around them where it means you can't switch from one provider to another. They're high, but they're not that high. Storing the data inside of a cloud versus sending it somewhere else, the transmission of it out of that cloud costs the same is storing the data for roughly four months. So it's high, but it's not egregiously so. There's a concept known as data gravity, because it's expensive to move data around, you keep the data where it is and then the compute workloads such as AI stuff starts centralizing around that. We're seeing an inversion of that when only some providers are doling out access to the kinds of compute that we need in the form of Nvidia GPUs. So you wind up effectively having to bring the mountain to Mohamed, so to speak. As a result of this, we are seeing people do massive data transfer projects from where they are to get it close to these things, and it's sort of turning a lot of this stuff upside down in a very topsy-turvy way." - Corey Quinn

- "Now, a second barrier we looked at were the costs and the effort that customers need to put in when they need to re-engineer an app and move it from one cloud to another, and that makes switching hard. But also some of the difficulties they have in connecting apps that are hosted on different clouds. And we thought that could make multi-cloud more difficult. " - Tania Van den Brande

Discounting structures

- "And finally, we pointed to a number of discounting structures in the industry that we thought create quite strong incentives for large customers to put all or most of their cloud needs with a single provider, and particularly make it unattractive for those large customers to split their cloud usage between larger and smaller cloud players." - Tania Van den Brande

Inability to compete

- "But we're also worried that some of these barriers create a risk that the cloud market might concentrate even further towards the market leaders, and particularly that they make it difficult for small cloud providers and challengers to really go after customers that are already established on AWS and Microsoft. And that worried us because we thought that makes it more difficult for those challengers to start building their customer base and to really start gaining that skill that you need in cloud to be a more direct competitor in the market, or at least threatened to become a more direct competitor." - Tania Van den Brande
- "But I think at this point it's obvious that we have if not a monopoly, the next thing to it. These cloud companies talk in the language of monopolists, which always touches on these ideas of, oh, it's a fight for survival for them. They could be out-innovated tomorrow by a startup in their garage. Well, yeah, if you give that startup $6 billion of funding for all of these AI training runs they'll need to do and the massive hiring binges and the specialized hardware, yeah, then maybe. But I kind of don't see it." - Corey Quinn
- "If you're a fully vertically integrated company and you have cloud models applications, it's very possible that someone develops an application or a kind of model that has some features that you think are terrific. And so you copy the idea, you integrate it into your own system, and that new startup is effectively out of luck because you have the scale and ability to operate that in every part of your ecosystem and roll it out to many, many customers because you're one of these major companies that is integrated across the whole stack and that you have all these different applications." - Ganesh Sitaraman