Tech Summit on Artificial Intelligence: A Quote Book

Data and Models Edition

April 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 second panel, we hosted the following panelists:

- **Cory Doctorow**, science fiction author, activist and journalist
- **Jonathan Frankle**, Chief Scientist (Neural Networks), Databricks
- **Amba Kak**, Executive Director, AI Now Institute
- **Stephanie Palazzolo**, Journalist, The Information, covering artificial intelligence

Panel Summary: Public reports indicate that certain AI foundation models involve hundreds of billions of distinct parameters that have been traced using many terabytes and trillions of tokens of data. The discussion underscored that the methods of data collection and model development have implications for both competition and consumer protection.

The panelists discussed that dominant firms have access to large amounts of public or private data through existing product lines and/or through changing terms of service. On the consumer protection front, large volumes of data are being used to train AI models, raising key questions for policy makers, such as: What are the legitimate business purposes for collecting, using, and retaining data? Are there types of data that should not be collected, used, or retained? Do consumers know how their data is being handled, and can they do anything about it? What happens if a company adopts more permissive data practices – for example, to start sharing consumers' data with third parties or using that data for AI training – and only informs consumers through a surreptitious retroactive amendment to its terms of service or privacy policy?[1] And what happens if companies misrepresent or don't fully disclose their privacy and confidentiality practices?

On the competition front, panelists expressed that incumbent tech firms have access to large amounts of data through the existing product lines, and that there are challenges associated with competing in AI development related to access to data, resources, and investment. This raises questions about whether the AI models will be developed and deployed in a way that fosters competition and introduces new competitive pressures on incumbents, or whether challenges associated with access to data and other resources might steer AI development in a direction that protects or enhances market power.

The FTC will need to remain vigilant in evaluating these issues as we pursue our joint competition and consumer protection mandate.

Why a Quote Book? The voices of people on the ground 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 – including the engineers building next-generation cloud-computing platforms, the data scientists training AI models, the venture capitalists who are funding new innovative startups or the startups building companies to improve consumers' lives – experience the effects of innovation in real-time.

The FTC recognizes that this is not a representative sample of all stakeholders, and we strive to continue to listen and engage with a variety of perspectives. The quote book aims solely 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.

[1] https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/02/ai-other-companies-quietly-changing-your-terms-service-could-be-unfair-or-deceptive

Themes

A Positive Vision for AI-enabled technologies

Governance grounded in human needs

- "It would mean having an information governance regime that thinks about information beyond being property. It thinks about information, and the harms that arise due to privacy invasions, to labor violations, to consumer rights violations." - Cory Doctorow
- "This current AI race is based on certain assumptions about both scale and speed as a proxy for progress. And it's a view that's based on narrow benchmarks, it's one that never really properly contends with the longer term environmental, or labor impacts, or the impacts on our information environment." - Amba Kak
- "[W]ho gets to decide and shape what counts as innovation, and what counts as innovation for the public good? I think that one way forward is to really go back to the drawing board, or the table, which is currently populated with VCs, big tech firms, and companies that they invest in, and really have a much more broad ranging conversation that is dominated by public, rather than very narrow private interests about what counts as innovation and what is innovation in the public good? And try to, I think, shape that trajectory more actively rather than be passive recipients or subjects of the tech trajectory." - Amba Kak

Ability for startups to compete

- "Success means leveling that playing field both within industry, and within academia, that more players have a chance to compete with the big industry labs." - Stephanie Palazzolo

Shared resources

- "I've always hoped that, I wish there were just centralized shared resources for improving the safety and quality of AI models. I wish there were a dataset that my students at Harvard, or my friend who just founded a company could pull off the shelf, train their model on that a little bit further, and know that they're getting something that is going to behave in ways that everybody would find to be at least a positive development, in terms of moving the model in the right direction." - Jonathan Frankle
- "I'd love to see in the public policy world, shared safety resources. Certainly something that a lot of organizations are talking about working together on, but having that publicly funded would go a very long way toward just raising the bar across the board, and ensuring that... or at least to minimizing the trade-offs that a company, the race to compete and the race to get cool things out there." - Jonathan Frankle

Flexible thinking on open sourcing

- "...thinking about open sourcing in a more flexible way. It really is treated as a binary right now, in terms of either you open source your model and throw it out there to the world, or you keep it secret and never tell anyone about it, or how you built it, or share it with anyone. I have to believe there are good middle grounds." - Jonathan Frankle

Potential consumer protection issues related to AI model development

Direct consumer harms

- "Although we described some very valuable things by calling them property, that the most valuable things in the world we describe with non-property language, and that's people. Harming someone is not theft of their integrity. Killing someone is not theft." - Cory Doctorow
- "People really do feel a lot of pressure to race to get things out there." – Jonathan Frankle
- "If it's possible to pay for the data, and still enact the same harms, still displace creative workers with the work that they've done for you, still possible to produce grotesque privacy invasions in the form of non-consensual pornography, still possible to harm people by mining their data to make inferences about them that are adverse to their interests, then we have managed to fail to solve the problem, while still creating a bunch of law, and wasting a bunch of time, and incidentally, also creating a regime in which people who have the money to pay for data licenses are the only people who get to play." - Cory Doctorow

Labor issues

- "...Giving creators more to bargain with is like giving bullied school kids extra lunch money. There just isn't an amount of lunch money that gives the bullies enough that they decide to just hand that over to the kid and get the kid fed. And a regime in which we say as a condition of training a model you must first license the content, is not a regime in which creative workers are defended from our creative employers." - Cory Doctorow

Lack of transparency creates risks to consumers

- "Another team at Google built BERT, which was an important early model. And the people at OpenAI scaled them up, and led us where we are today. But especially in the past couple of years, that's really shut down. Because now, it's competitive intelligence. And nobody wants to give that away. We know a lot less about the inner workings of current generation systems than we do about their forebears. And that is one consequence of the intense competition that we're seeing." - Jonathan Frankle
- "My understanding from lawyers is nobody really knows how anything is going to turn out the first time around. So there's incentive for organizations to be pretty secretive about what data they're using, and how they're building models, because it reduces risk in a really uncertain environment. And even for some of the most popular freely available models like Llama2 and Mistral, we can access the weights of the models. We [still] don't know how they were built, [though]." - Jonathan Frankle
- "We are seeing that trend towards more restrictions on publicly available data, higher cost of acquisition, and a turn towards more non-transparency and opacity around who is even using what data." - Amba Kak

Potential concentration issues related to AI model development

Outsized power and bottlenecks

- "There is something intrinsically offensive about the idea that they'll take the product of your work and use it to make sure that they don't have to pay you anymore. I think that it neglects something very important about the structure of the creative labor market, which is that it's a monopsony in

which a small number of firms have enormous amount of bargaining power over the creative workers who generate the value for them." - Cory Doctorow

- "So I think a question for the very near horizon is to try to watch for how large tech firms leverage their existing relationships, and importantly, their skewed power dynamics with publishers, and with the media industry to maximize both access, but also push for exclusivity." - Amba Kak
- "Venture capital, and the tech giants have a very large role in picking what startups are going to win and lose, coming out of this AI boom. As I mentioned earlier, for instance, just the fact that OpenAI exists is stopping investors from backing certain types of companies. And I think for me, success means leveling that playing field both within industry, and within academia, that more players have a chance to compete with the big industry labs." - Stephanie Palazzolo

Outsized costs to compete

- "I remember talking to one startup in the weeks following that where that was their entire concept leading up to that event. And now they're kind of just like, "What are we supposed to do? We raise money on this idea, but now OpenAI, which has $13 billion of funding is going after the same thing. How are we supposed to compete against that?" So I think a lot of investors are just wary of backing startups that could even be in an adjacent area to what OpenAI is doing now because they're worrying that if OpenAI or Google or some other big tech giant goes up and tries that as well, that their investment just won't be able to survive." - Stephanie Palazzolo
- "It takes so much capital for [startups] to buy chips, to buy data, to hire people, it's actually much harder for them to generate cash and they tend to have lower margins than traditional software businesses that we've seen in the past. So this could obviously change as we move forward and as chips get more efficient, for instance. But part of me does wonder if VCs might come to kind of regret their actions of funding a number of these startups at very insane prices." - Stephanie Palazzolo

Barriers to entry (e.g. high-quality data, venture capital funding)

- "One of the most important expensive inputs to building these models isn't computed. It's the data hand labeled by humans to make the model good at specific things. And for any of the fledgling startups that are out there, that three to six months, and those millions of dollars probably have to be weighed against racing to market, and making a name for yourself in a really competitive environment. And that incentivizes risk-taking." – Jonathan Frankle
- "…Quality data, and this has everything to do with labor, data sets with high levels of human curation, and feedback, and niche data sets, especially in high impact sectors like healthcare or finance, data sets that come with assurances of accuracy, and legitimacy, and diversity at scale, these are all becoming a very key source of competitive advantage, especially in the hyper-competitive generative AI market that Jonathan was also just describing." - Amba Kak
- "On one hand you have a lot of companies that are getting tons of venture capital at really insane prices, but on the other hand you do have a number of startups that are really struggling to find any funding at all." - Stephanie Palazzolo
- "OpenAI, even though it is in many ways still just a startup, it has kind of fallen into this role as a market leader and it's indirectly kind of choosing which startups win or lose in this AI wave. So I've noticed a lot of VCs are really hesitant to back things that either directly compete with OpenAI or even are in areas that OpenAI could maybe go into at some point." - Stephanie Palazzolo
- "Venture capital, and the tech giants have a very large role in picking what startups are going to win and lose, coming out of this AI boom." - Stephanie Palazzolo
- "So I think a lot of investors are just wary of backing startups that could even be in an adjacent area to what OpenAI is doing now because they're worrying that if OpenAI or Google or some other big

tech giant goes up and tries that as well, that their investment just won't be able to survive." - Stephanie Palazzolo

Misaligned incentives

- "Having the money to pay for data licenses is not correlated strongly with being someone who will not harm the public, that there are lots of incumbents, with lots of money, who've got a strong track record for being the last people we want to lead us into the future." - Cory Doctorow
- "I think the pitch to investors is that we are going to tell hospitals you can fire half your radiologists and double their output, and that is not the AI productivity and benefit world that we want. That is the alignment problem that we're worried about." - Cory Doctorow
- "...there is nothing that is inevitable about the current trajectory of AI. That is really important to keep remembering and reminding everybody. Because I think this current AI race is based on certain assumptions about both scale and speed as a proxy for progress. And it's a view that's based on narrow benchmarks, it's one that never really properly contends with the longer term environmental, or labor impacts, or the impacts on our information environment." - Amba Kak
- "We are nowhere near a place where a bot is going to steal your job, but we are well beyond the point where your boss can be suckered into firing you and replacing you with a bot that fails at doing your job. And I think that's the real AI alignment problem we should be thinking about." - Cory Doctorow

Access to talent and capital

- "I've noticed a lot of investors going after companies that are founded by ex-OpenAI researchers or maybe scientists that were at Google or from very... some of the top colleges in the U.S. And that kind of makes it harder for founders that maybe come from other types of backgrounds to get funding and to get capital from these investors." - Stephanie Palazzolo

Data advantages

- "Big tech firms have a very clear advantage here from the last decade of commercial surveillance." – Amba Kak
- "As a related point, we also don't know if and to what extent these data advantages will port to the so-called AI startups that they are strategically investing in, potentially creating new forms of dependency and power asymmetries outside of those that already exist via the compute and cloud arrangements." - Amba Kak
- "I think one of the biggest misconceptions I typically come across in our field, and especially when I chat with folks in policy is everyone just assumes that because they know of OpenAI and ChatGPT, that's the only business model and that's the only way of operating." - Jonathan Frankle
- "Even getting the data to do this is incredibly expensive. One of the most important expensive inputs to building these models isn't compute. It's the data hand labeled by humans to make the model good at specific things. And for any of the fledgling startups that are out there, that three to six months, and those millions of dollars probably have to be weighed against racing to market, and making a name for yourself in a really competitive environment. And that incentivizes risk-taking. So that's certainly one of the consequences of competition that I imagine is on the minds of a lot of people I know at small startups." - Jonathan Frankle

Inability to sustain a competitive foothold

- "'How are these businesses sustainable in the long run?' And the question of business model, and how startups that are making open source models make money is very important if you're one of their customers that's depending on them to build a product on top of their open source model. And

I guess one question I have is, this is a question that's come up a lot with other types of open source software and tech as well, which is, 'How are they going to sell this?'" - Stephanie Palazzolo

Open models

Framework

- "So I prefer to take apart the term open source when it comes to models, and think of it as two things. It's about **access to models, and transparency** about models. So by access, I mean, 'Do you have access to the model weights itself, and not just a way to talk to it? Can you literally take the model, and manipulate it, and work with it yourself?' And the other is **transparency**. 'Do you know how the model was built?' To give you an example, we don't know a lot about how Llama 2 was built. The paper doesn't say a ton about the data that was used to train the model, or the details of the hyperparameters." - Jonathan Frankle

Benefits

- "With an open-source model, you control your own fate. You also get greater control of your cost structure. You can serve the model yourself, and you know exactly all the inputs that are going into that down to the cloud level. But I do want to emphasize. These are all benefits of access to models. I haven't said anything about the benefits of transparency. I think the transparency part is just a little more complicated." - Jonathan Frankle

Risks

- "Open source AI should not be assumed to be a stand-in or a substitute for structural interventions across the AI stack, because those firms are also vulnerable to the same dependencies, and will also need the same protections from practices like self-preferencing that I heard were discussed in the previous morning panel on compute as well." - Amba Kak
- "Open source companies are still operating in a highly concentrated ecosystem in which the largest firms retain both the resource, and data advantages, and network effects." - Amba Kak

Open questions

- "I personally don't fully understand the business model behind this. And it may just be that I'm an AI researcher, and not a business person. And maybe I'm missing something. But I don't know if it's sustainable. I don't know how you justify doing this day in and day out. So I don't know whether if I were a business relying on the open source models today, I don't know if you could rely on that as your long-term strategy right now." - Jonathan Frankle
- "How are they going to make a product on it that people are willing to pay money for?" - Stephanie Palazzolo

Looking forward: How to approach AI model development

Data minimization

- "Data minimization is a very core part of the toolkit. It's a principle that needs to be defended against a I would say, an existential threat, which is the idea that AI innovation requires a no-holds-barred regulatory orientation to data, that regulation is at odds with innovation." - Amba Kak
- "Data minimization as the key principle that has been endorsed and enforced over a decade, and just to say that this principle is more important, not less in the age of AI. Incentives already exist for invasive and irresponsible commercial surveillance. But this current version of the AI race definitely pours gasoline on that problem." - Amba Kak

Brightline rules

- "Data minimization isn't new at the highest level. It's been around for more than a decade globally in various forms, in various laws including the GDPR. So I guess, 'Why hasn't it worked,' I think. Or, 'Why hasn't it prevented some of these, the worst privacy invasive practices?' I think there, the lesson, if anything, is one on not allowing too much room for interpretation. Because I think where maybe the first decade of data minimization came to a head was on the question of, 'Is behavioral advertising a legitimate business purpose? And if it is, does that mean we can just maximize, collect as much data, and keep it forever?' And I think as we look forward, acknowledging those administrability challenges, acknowledging that an interpretive wiggle room will be abused, to really focus on bright-line rules that don't allow that, that make it very explicit that AI training is not a free card to break down all your data silos, to violate purpose limitation, that we want to draw bright-lines around restricting particularly the use of the most sensitive data like biometrics or related sensitive attributes." - Amba Kak

Integrate competition and consumer protection

- "[P]ushing for more integrated regulatory approaches that don't silo out the consumer protection side of things, and the competition side of things. Because we have, again, seen how some of the largest firms really took advantage of that over the last decade to amass the information asymmetries that they have, and further concentrate their power." - Amba Kak

Strengthen laws

- "To address these harms, we have to reach to things like labor and privacy law, not copyright law. It is not enough to merely have the right to feel affronted by conduct of firms. We should have the right to do something about it." - Cory Doctorow
- "I think Americans often underestimate just how primitive the state of American privacy law is. The last time we got a really big muscular improvement to our national federal privacy regime was in 1988, when Congress got worried about video store clerks leaking their video store rental history, and passed a law prohibiting that activity. The Internet's come a long way since then." - Cory Doctorow
- "If you are worried that TikTok is making millennials quote Osama bin Laden, or if you're worried that Facebook made your Grandpa a conspiratorialist, or that Instagram is making your kid anorexic, or that protesters at Black Lives Matter demonstrations, or the people who attended the January 6th riots are all being identified by Google through reverse warrants, you are someone who cares about privacy, as is anyone who is worried about the privacy implications of AI, whether that is models memorizing, and then regurgitating private information as we've seen, where sensitive information from medical histories, or resumes, or commercial databases of purchase history are sometimes being memorized and coughed up by these models, or whether you're worried about the truly grotesque generative AI problems with image and video generators, things like non-consensual pornography generation, copyright's not a great tool for dealing with this. But privacy law would certainly give you an awful lot of remedies to deal with." - Cory Doctorow

Regulation

- "...in a market where, to Cory's point, the business model is entirely elusive, the regulators and the public I think need to keep a close tab on where this is headed, if all roads do lead back to some data-driven personalized advertising paradigm." - Amba Kak
- "we need to start getting quite specific about drawing guardrails, and drawing them quickly, that we're not left cleaning up the mess after those incentives have already been supercharged." - Amba Kak

Integrating disciplines for decision-making

- "I think there's just so much room for professors and researchers at these universities to be part of this conversation. And they don't have the same incentives that profit driven companies do. And I think we really need to be encouraging that a lot more."
  - Stephanie Palazzolo
- "It's not open versus closed source. It's not more laws versus letting people innovate. It's not the academic side versus industry. There's a lot of gray area here, as we all talked about today." - Stephanie Palazzolo

Grounding rights central to human needs

- "Merely giving tradable property rights is always going to be inadequate in the same way that we don't solve the problems of a lack of organ donors by creating property rights in kidneys, and then just letting people sell their kidneys. We need rights that deal with our information in a way that is cognizant of, and sufficiently important that we recognize its gravity, and its centrality." - Cory Doctorow
- "We need to think about the problems of data beyond a property rights regime, beyond the idea that if you make data, it's your property. And someone else has to pay you, and get your permission before you use it. Because what we want to make sure of is not that everything in the models is paid for, but that the public and other stakeholders aren't harmed." - Cory Doctorow

Improved media framing

- "It's the responsibility of, especially the media, and people like me, even though it's much more easier to write stories, and just say, 'Oh. It's X versus Y,' I think it's up to all of us to make sure that we're discussing this, keeping those gray areas and nuances in mind." - Stephanie Palazzolo