The Economic Club of New York

ESTABLISHED 1907

The Economic Club of New York

116th Year
728th Meeting

Mustafa Suleyman
Co-Founder and Chief Executive Officer
Inflection AI

October 12, 2023

In-Person/Hybrid Event

Moderator: Marie-Josée Kravis
Chair, Museum of Modern Art
ECNY Chair Emerita

Barbara Van Allen, President

Introduction

Good morning and welcome to the 728th meeting of The Economic Club of New York. I'm Barbara Van Allen, President and CEO of the Club. The Club is known as the nation's leading nonpartisan platform for discussions on economic, social, and political issues. And we've had more than 1,000 prominent guests appear before the Club over the last century.

I'd like to extend a warm welcome to students from the CUNY Graduate Center, Rutgers University, and Mercy University joining us virtually today, as well as members of our largest-ever Class of ECNY Fellows – a select group of diverse, rising, next-gen business thought leaders. As a reminder, applications for the 2024 Fellows Program are now available online.

Today, as part of the Club's Author Series, I'm honored to welcome our special guest, Mustafa Suleyman. Mustafa is a serial tech entrepreneur and Co-Founder and CEO of Inflection AI, an AI-first company redefining the relationship between humans and computers.

Mustafa previously worked at Google as Vice President of AI Products and AI Policy.

Before that, he co-founded DeepMind, which was bought by Google in 2014. As Head of Applied AI, he contributed to the team's major successes in AI research and applications for over 10 years. His book, *The Coming Wave: Technology, Power and the 21st Century's Greatest Dilemma,* was published last month. You all have copies on your chairs. And this was described as an "excellent guide for navigating unprecedented times" by Bill Gates.

Mustafa's accomplishments have been recognized both in the United States and the U.K. In 2019, he received the Commander of the Most Excellent Order of the British Empire for his influence in the UK technology sector. The same year, he accepted the Silicon Valley Visionary Award.

He's a Senior Fellow at The Belfer Center for Science and International Affairs at the Harvard Kennedy School working on the geostrategic challenges of future AI systems. He's also a member of the Board of *The Economist* and a member of the Steering Committee of the WEF's AI Governance Alliance.

The format today will be a conversation, and we're honored to have Marie-Josée Kravis, ECNY's Chair Emerita and current Chair of the Museum of Modern Art, as our moderator. As a reminder, the conversation is on the record. We do have media, as I mentioned, online and in the room. Unfortunately, he has a very tight schedule so he will not be able to sign books afterwards as he needs to get on to the next meeting. But thank you. With no further ado, if you'll all join me in welcoming both Mustafa and Marie-Josée. At this time, they will take the stage.

Conversation with Mustafa Suleyman

CHAIR EMERITA MARIE-JOSÉE KRAVIS: Well, thank you, Barbara. And thank you everyone for being here, and especially thank you, Mustafa, for doing this, because I know that you've been on the road with your book, but you're also running a company.

MUSTAFA SULEYMAN: That's true.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: It's really a great honor and a privilege to have you here. Thank you. And I thought what we'd do today, and I'm going to incorporate questions that were submitted by members in our conversation, so I'll take into account many of your suggestions for those who are here and those who are online.

Let's maybe go back to how this all started, how your interest in technology evolved, and then we can talk a little bit about DeepMind, about your book, and we'll talk about Inflection AI. So let's start at young Mustafa, with a philosophy major, or maybe even before. How did this interest in technology develop?

MUSTAFA SULEYMAN: Yes, I've always had a bias towards action and doing things. So when I arrived at Oxford, I studied philosophy. I was immediately sort of quite frustrated. I was like, oh, my God, what did I get myself into? This is very distant from getting things done. And I had a sort of equal bias for basically doing good. I was very principled. I had become an atheist. I grew up as a very strict Muslim. And thankfully, when I got to Oxford, I became an atheist quite quickly and had some really incredible moments studying human rights with a few of my professors.

And I guess that combination of a bias towards action but also being very deeply principled led me to want to drop out and start a charity, which I ran for three years and it's still going 20 years later. It's a telephone counseling service. And at the same time I also started a business while I was at Oxford, my first business, which was an electronic point of sale system back in 2003, where we sort of integrated like little PDAs and Wi-Fi systems into, or sort of basically giving an internet plus a PDA system to a restaurant. And it was very unsuccessful. We lasted a full year. It was brutally painful. I lost money for my investors. And it was a great lesson. I immediately started another business, this time while starting the charity, this time selling fruit juices and smoothies.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: And that was quite successful.

MUSTAFA SULEYMAN: That was quite successful, yes. So yes, I mean, look, I'm a very abstract thinker. And I am very good at sort of challenging and questioning and asking and being just sort of unabashedly ashamed of just trying to get to the bottom of something, and that has served incredibly well. Throughout my, sort of career, I've been in interlocutor essentially, a translator, sort of trying to distill and synthesize complex ideas for my peers and my colleagues and my teams.

And that has actually turned out to be very valuable for understanding machine learning. I mean machine learning is a very, is much more accessible than I think people realize. Set aside the software implementation, the code itself, you know, it takes practice and requires careful deliberation and adherence to best practice and so on. But just, I'm sure many people here are economists, I mean it requires some level of abstract thinking about variables that intersect with one another, the change over time with respect to various other considerations. And that is a very similar framework to how machine learning works. Machine learning is really just mathematics that can be expressed in a page or less. And that can give you a pretty good introduction for what the system is trying to do.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: Well, you're trying to make it rather simple and approachable, but take us then to the leap that took you to DeepMind because that was a real breakthrough.

MUSTAFA SULEYMAN: When we started DeepMind, so my co-founder at DeepMind, Demis Hassabis, and our other co-founder Shane Legg, were finishing their post-doctoral work at UCL, at the Gatsby Computational Neuroscience Unit. I'd known Demis for quite a long time and we were both interested in how to change the world. I mean we were obscure and strange people in the sense that like we were always thinking about very long-term effects, and what would it take to make an intervention, or to take a big bet.

And there weren't very many people who were thinking about a 10 or 20-year time horizon. I was thinking about that because I'd become very interested in climate change and I was working as a negotiator at the Climate Negotiations in Copenhagen in 2009. Very frustrating experience, mostly a failure again to reach consensus. And I became so frustrated that I was like, really what we need here is better science and technology. We need to invent our way out of this problem. We're not going to be able to talk our way out of the problem.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: And, of course, Demis was very interested in games.

MUSTAFA SULEYMAN: Yes.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: So, is that what led you to first chess and then Go?

MUSTAFA SULEYMAN: Exactly, yes. We were interested in how could we model an environment well enough to understand all of its mechanics, make predictions in that environment, and then use it to intervene in the course of that environment? And that's essentially what we're trying to do in climate change. It's what we were trying to do in games. Games is really just a simplified representation of the real world. I mean just like the rational market hypothesis, sometimes a dangerously simplified representation, but a useful one for modeling and making progress. And that was why we picked games, because at the time, in 2010, when we started DeepMind, you know, the algorithms just weren't good enough. And so we had to simplify the world in order to mirror the complexity or capability of the algorithm.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: But you described in your book how even you, while watching these games unfold, you, I mean in the plural sense with Demis and your team, how you even surprised yourselves as to how successful the model had become.

MUSTAFA SULEYMAN: Yes, I mean I think people often sort of, we were just saying earlier, like people look back at the last year or two and frame it as a kind of bolt from the blue, you know, a surprise that's come out of nowhere. And yet I definitely think it's possible to look back over the last decade and clearly see very many important moments where the models have done something surprising that have given us confidence that we're actually on a trajectory of steady improvement and not accidental sort of explosive progress.

So the very first moment that I experienced that was when we saw, we trained in AI to learn to play the old-school Atari games, 50 or so Atari games to human level performance. And when we say trained, the crucial distinction is that we didn't handcraft a set of rules that say if you're in this kind of position, take this kind of action. We simply designed a very simple reinforcement learning algorithm that says take the score that you're accidentally stumbling into at any given moment and associate that score with a set of actions that you randomly took in the frames in the runup to that successful moment.

And so this associative learning or reinforcement learning actually produced with many, many sort of millions of iterations of self-play pretty complex behavior. You know, in the case of Space Invaders, the DQN algorithm at the time could dodge incoming enemies. It could like shoot missiles and quite precisely hit targets. And then in order to play the game Breakout, which was where you control a paddle at the bottom and bounce a ball up and down, it developed a pretty cool tumbling strategy that funneled the ball up the side by targeting the same block continuously and then bouncing the ball up and down at the top to get maximum points with minimum effort. It's a very simple idea. Most human players, if you're addicted enough, discover it.

But the crucial thing is that this was new knowledge that wasn't programmed into the system, that it had learned through self-play and through a very simple reinforcement learning algorithm that could be expressed in a single sentence. And that was back in 2012, published in 2013. And we essentially took the same high-level framework and repeatedly applied it many, many different domains over the course of a decade. And, you know, with a few tweaks, that's pretty much where we've ended up right now.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: Well, yes, the games, but you did achieve a major breakthrough, especially as a scientific tool, with AlphaFold, because that maybe follows the same approach, but is a real major breakthrough in terms of enabling scientific research. And, in fact, was just awarded the Lasker Prize here in New York I think last month.

MUSTAFA SULEYMAN: Yes, last month. Yes, I mean the cool thing about it is that our thesis was that we could take the same basic algorithm and apply it to lots of different domains. So before we started working on protein folding, we actually published three nature papers over the course of three years, the first work in OCT scans, diagnosing

50 blinding conditions to consultant ophthalmologists performance, including not just the diagnosis but a treatment, a proposed treatment pathway. We did the same for mammograms, chest X-rays. We did the same for the prediction of acute kidney injury and sepsis, which are the two biggest killers, avoidable killers if you're in admission in a hospital.

And, you know, with a little bit of hand-waving, just to cut it short, the same basic algorithm was applied in these really quite different modalities. And now just in the last year, other researchers have published incredible results showing that just from, you know, these three-dimensional OCT eye scans, you can tell all kinds of interesting things like propensity for cardiac arrest, Alzheimer's, diabetes, glaucoma, even gender, age, and to some extent, race actually. So it's very interesting. These models are picking up on signal in the raw data that we, as humans, have no chance of being able to see. But because they have been able to experience millions of cases, they can associate patterns that are basically beyond us.

And so the problem of protein folding that you referenced, AlphaFold, is actually really very similar. From a small number of data points, we're trying to extrapolate that actually this would be a useful and accurate prediction of how a protein unfolds. And then ultimately what we're really interested in and the primary motivation of that project was how could we map the functional structure to the real world assays. Like what does this thing do and how can we learn from this folding that we can then predict that has these functional properties? That's the real breakthrough and that's yet to come. If we knew that, I mean that really is completely transformative. We would be able to manipulate biology to design really perfect drugs, to design new compounds and so on and so forth.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: And are we anywhere close?

MUSTAFA SULEYMAN: I don't know. I don't think so. I think that is really, really hard because we don't have enough real-world data about functional assays mapped to molecular structure yet. So it's probably, I think, not within the next five years, maybe ten years. It's certainly within a 20-year period. So I think we should, that's what said in my book, there's so many reasons to be optimistic. I mean we really are heading towards an era within our lifetimes, in the next 30 years, of radical abundance. I mean we will be able to produce more with less on a completely unprecedented scale. Things are going to get a lot cheaper.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: Well, let's get to your book because that's probably the, I found, one of the most interesting features of your book is that you don't deal with one technology. I know there's been, because of large language models and so on, and ChatGPT, so much interest in large language models. But you go beyond that and deal not only with AI but with this convergence of technologies – synthetic biology, AI, fusion, genetic engineering. And it's this wave that you describe, and you probably can describe it, I'm sure you can describe it more eloquently than I can. But it's this convergence that defines this point of inflection that we seem to have reached.

MUSTAFA SULEYMAN: That's right. And I think that it's the same underlying trend, which is computational power combined with vast amounts of information allows us to reveal underlying patterns in arbitrary data. That's, I think, the key general purpose. It's almost like the meta technology in itself. I mean intelligence, it's hard to sort of frame intelligence alongside all of the other general-purpose waves that we've experienced in the history of our species because intelligence is the ultimate technology. It's almost like a meta technology. The ability to use tools is a way to amplify our own capabilities but also turbo-charge these other waves that are on their way like you said. And, you know, I think that that is unlike anything we've ever seen before and likely to produce value like we've never seen before.

I mean imagine having access to research assistants and creators as capable as any human that has ever lived and being able to direct those sort of artificial intelligences, if you like, to specific purposes. And I think it's important to be clear, that doesn't necessarily mean those beings have autonomy or that they would be conscious or sentient in any way.

If we engineer this correctly, it's possible to take the good things about being human – our ability to make predictions quite accurately, our ability to synthesize vastly different types of information, and our ability to creatively put those things together to invent new things – without having the emotion, the biases, you know, all the kind of weak judgments, the value judgments that come with those things, and certainly without having autonomy or sentience. And so in that sense, I see it really as a tool to turbo charge us.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: So, some of your colleagues – I'm thinking of Yann LeCun or Yoshua Bengio and so on – have in a way disparaged large language models and said that they're more or less a distraction in the path towards artificial intelligence. Do you share that view?

MUSTAFA SULEYMAN: I think a healthy academic debate is essential and there is a risk that we are converging around a single architecture, you know, to the detriment of all others. And so a lot of academic departments are frustrated that they're not getting well-funded unless they focus singularly on large language models. And that's a concern.

At the same time, you know, empirically speaking, this is working incredibly well. We've sort of been through this first phase of classification where the models understand raw data and can label it. We’re now in this phase of generation, where they understand the underlying data well enough that they can produce an accurate prediction of what might come next, be it in text or images. The third wave or phase is going to be interaction. So you can use that generative AI to get feedback from other AIs or from other humans in the real world. And the next phase, which we’re just scratching at in parallel is going to be multi-modality. Not just with images and video, but any sensor data.

And so in principle, the only limit to being able to integrate those other modalities, I mean think of every ray of light, every element of sound you could possibly imagine, all the sort of tactile data that comes from physical interactions with the world, all of the arbitrary time series data that we have from so many other sources that when you just look at the raw form means basically nothing to a human eye. So, you know, these models in a quite unsupervised way, without hand-curating any of the inputs, can actually pass very multi-modal data from very different sources, not just sort of audio, text, and image, but these other time series data.

So I think that seems to me almost certainly the case that we’re going to be able to integrate all of that and make use of it. So I’m not persuaded that this isn’t a good route to continue pursuing even though...the problem, I think, with some academics is that there’s a desire to have this like very purist interpretation. You know, my view of the world is being broken but as good scientists they should, and they do, I mean they observe...

CHAIR EMERITA MARIE-JOSÉE KRAVIS: The perfect is the enemy of the good.

MUSTAFA SULEYMAN: Yes, yes, exactly. I mean it's working perfectly well.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: So when you say it's working, and your book is very balanced in terms of the positive impacts but also the more dire impacts. Can you maybe jump to those? Because you've alluded to the multi-modality. So the risk of disinformation, of misinformation, the risk of cyber-attacks, we can think of more nefarious uses also of AI as creating pathogens and disseminating pathogens and so on. How worried are you about the more negative effects? And we'll get to the positive...

MUSTAFA SULEYMAN: I'm very worried about it, and I think it's right to be concerned. I think that the better framing is, rather than assuming that we're going to be stuck with the issues of bias or the issues of hallucinations, you know, for the next decade, assume that those issues are going to be dealt with very soon. And probably in the next two years, I think we're going to largely eliminate hallucinations entirely. And I think that the models are going to get so controllable that they will have a bias to the extent that you shape the model to have the bias. I mean they're going to have very few unintended biases. I mean we're already seeing that.

So it's worth remembering that each generation of these models is ten times larger in terms of computational consumption than the previous model. So when we talk about going from a frontier model of generation 2, like a GPT2 or GPT3, GPT4, the difference between 3 and 4 is 10X. The difference between 2 and 4 is 100X compute. I mean we really haven't, we don't see those curves very often in any area of science and engineering. And the fact that we can now predict with very high certainty that over the next five years, models will be trained that are four or five orders of magnitude larger than they currently are today at GPT4. You know, 10,000 times or more larger.

We can also then extrapolate what capabilities might arise with more computational power. Because the role of the computation is essentially to attend to, pay attention to more of the underlying training data with respect to itself. So you have these huge, huge databases of tokens, trillions of tokens, trillions of words, and the models learn the connections between all of the different words. And if you have a small amount of computation, then the model can only use that computational budget to attend to so many of the possible relations between all the different words. If you have essentially infinite, then in theory it can attend to all, it can make a wall-to-wall connection between all of the underlying training data.

And what we saw in GPT2 was a model that was essentially incoherent. It could barely complete sentences, so, like unrecognizable. And now with GPT4 it's approaching human level performance at language generation across a wide range of tasks. And it's not just the accuracy that has increased. It's the controllability. So with GPT4, even compared to GPT3.5, it's much easier to provide a specific set of instructions and see a specific set of generations more closely aligned with that behavior policy.

You know, at Inflection, my new company, our first model, Inflection-1, beats GPT3.5 on all of the academic benchmarks, including PaLM, Google's model, and LLaMA, the open-source Facebook model, or Meta model. By the end of the year, we will have beaten GPT4 on every benchmark and we'll be the best model in the world, depending on what Google does with its new model. They might publish before us.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: And what GPT5 brings next spring.

MUSTAFA SULEYMAN: GPT5 will be, you know, we will produce a GPT5 sometime in the spring, basically being 10X larger than the previous model with our own proprietary modifications to how it performs. So, to me, it's quite predictable that the models will get much more accurate and much more controllable.

The second sort of big step forward we're likely to make is that instead of producing a sort of one-shot answer to a question, like at the moment you have to ask a question to an AI and it gives an answer. And that's called one-shot. Instead, you're going to give it a more abstract goal, you know, like a high-level goal and it's going to go off and produce a string of actions that are all consistent with one another over time. And those individual actions might be, you know, produce a piece of text, then go and produce an image, then go and generate an email, and then send that email along with the text and the image to a human, pass their input, you know, take their feedback, integrate that back into the development of the text and the image and send again.

And I think that's quite important because what I've abstractly described there is a very fundamental capability that all of us use in our everyday work. And these models are likely to be able to do that in the next two orders of magnitude, let's say, I mean I could be more aggressive, but roughly on that order.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: But it's interesting. I mean you speak about that and you speak about the spring, but we in the U.S., in the fall, will be in an election, well, all of next year, but culminating in the fall, in an election period, and there is quite a bit of anxiety as to how these models might be used or misused in terms of influencing political information. Are you worried about that?

MUSTAFA SULEYMAN: Yes, I mean it is going to be possible to very cheaply produce, and I think the cost is the other consideration, is that the, you know, as we've seen with the history of all of our technologies, as long as humans have been alive, if it's valuable, then it gets cheaper because everybody wants access and we find more efficient ways of doing it. And so we should expect these AI models to be available in open source, almost at the cutting edge of frontier performance. Maybe a couple of years behind the absolute frontier.

And everybody will be able to get access, as we've seen with the open-source diffusion models for image generation. It's quite an incredible thought. I mean every image that is available on the open web has now been compressed into a two-gigabyte file that you can download and run on a good Mac, a good laptop, and it can generate novel images that you've never seen before. That's really quite remarkable.

And that's kind of what I mean by the plummeting cost of power, because if that trajectory holds, and obviously I'm predicting that it is, then it won't just be the ability to generate images, it will be the ability to generate sequences of actions over time that will be compressed into a very small transferable unit of power and usable by anybody. So the synthetic misinformation thing around the elections, sure, important, we should pay attention to it, but I think that's kind of a bit like the discussion we've been having the last few years around bias and around hallucinations.

Important questions, don't get me wrong, I really don't mean to trivialize them. I just think, step back and consider the broader context over a five and ten-year period. You know, we are really going to have extremely capable intelligences that are approaching human level performance as amazing project managers, assistants, inventors, you know, kind of scientists really, certainly entrepreneurs. And I think over a ten-year period, those are going to be cheap and widely available in the open-source. And I think that's the structural political question to focus on. Because what does it mean that anybody can wield state-like powers.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: Well, I was going to get to that point, is that because of the ubiquity and availability of these technologies, it's very, very different. People often talk about proliferation and non-proliferation or they talk about containment. And you raise that question, is containment even possible when these technologies are available to non-state actors, good and bad?

MUSTAFA SULEYMAN: It's worth saying that, you know, we have never contained something with these characteristics. We've contained other types of power before. You know, in fact, most new technologies that we invent are, by default, contained, right? So, aircrafts, for example. You have to get a license. Every component is regulated to within an inch of its life. They can only travel on certain tracks. There are regulations on every part of it, right? The same with cars, the same with handling certain sensitive biological chemical substances. I mean we have very effective regulatory frameworks, which protect the public interest and minimize harms and maximize benefits. But they take time to unfold. And so far, especially with nuclear, you know, non-proliferation, the structural dynamics of those kinds of problems are very different to these diminishing cost curves and this kind of race towards zero marginal cost.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: But also the limited number of countries that had nuclear weapons, and we could discover and there could emerge some rogue actors, but nevertheless it is a limited number of participants in these non-proliferation discussions. How can you do that when a technology is so widespread?

MUSTAFA SULEYMAN: Yes, that's the challenge. So the good news about that is that the knowledge and know-how to be able to build a nuclear weapon is actually widely available. With a little bit of intent, you can even find it online. The restriction is it costs a lot. It requires huge physical infrastructure, which is very observable. And clearly uranium-235 is scarce and hard to handle. And crucially, there was a taboo around the weapons. You know, it's a very significant fact that those two bombs were dropped, and they served as this eternal reminder to everybody of the horrific nature of these weapons. So whilst they are dual-use and it's clear that they're very valuable for power, the threat was also extremely well-articulated by history.

In this case, I think it's very hard to say that GPT4 has caused any significant harm. There's been a couple of cases and so on. But net-net, this is the most unbelievable benefit we could possibly imagine. And so the harm isn't going to arise in the traditional sense. There isn't going to be a sort of massacre in that sense. I mean in synthetic biology, it's easier to observe the harms because it is going to reduce the barrier to entry to manufacturing, you know, potentially sort of more transmissible, more lethal, engineered pathogens. I mean it is going to get easier to do that because you're going to have a coach that sort of teaches you if you don't have that training.

And I think the good news is that the open-source efforts, the big companies have been working very hard to prevent those capabilities. But, of course, some people might in the future, you know, train their own models that are deliberated designed to do that. And so you have to separate bad actors deliberately intending to do harm. And on the whole, the internet has been pretty well-policed. You know, if you think about 20, 30 years ago, how much harm could have arisen that hasn't been because it is actually well-policed, both by the big tech platforms but also by ICAM itself and governments.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: But I mean, some, and more and more groups would argue that we don't really understand the impacts that, for example, social media have had on children, on their education, on anxiety, envy, and so on, the rate of suicide amongst young girls, for example, in the 11- to 14-year category, influenced by Instagram or whatever. We don't really understand what the impacts have been. So it's too early really also to say that you'll have good actors and bad actors and that there will be a way of monitoring their actions. Don't you think?

MUSTAFA SULEYMAN: Yes, that is very true. I mean there are unintended side effects, which we're only now beginning to quantify as things go into production.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: Over time also...

MUSTAFA SULEYMAN: Yes, and that's one thing I worry about is that, you know, suddenly people will train models to be manipulative and be hyper-persuasive and try and participate in the electoral process. And we're going to have to confront a lot of sticky truths. People will – and are – want to use these models for romantic relationships, for sexual relationships, and in many cases they already are. There are millions of people having full-time relationships with AIs today. And it's the kind of reality that we have to confront. There are certainly going to be people that deliberately design these things for highly addictive, outrage-inducing, you know, engagement only kind of mechanics, and that is going to be concerning, super scary.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: And in terms of their development, the development of the technologies, do you think we're moving to a bipolar world where you have two AI superpowers – China and the U.S.? Or is that a model of the past?

MUSTAFA SULEYMAN: It's a model of the past and that's partly why I think we're moving towards it. I think the reality is that we have control of the primary commodity in this business. We, in the West, have Nvidia, and that is going to make a huge, huge difference. The export controls were essentially a declaration of economic war on China, which are holding them back from developing frontier general purpose models in a very significant way. I don't think it's going to be possible for them to train a GPT5. I think they're already struggling to train a GPT4. And so it's a major intervention into, you know, as they would see it, the development of their civilization. And it is as grand as that, I think.

I think sometimes we are not fully cognizant. When I talk to people about what the export controls mean, I feel that there's a kind of lack of empathy about how significant an intervention it is into their development trajectory. And maybe it's politically justified or not, I mean, I don't want to comment on that. But, you know, it's a big deal. It's a really big deal. They're not going to be able to catch up with us. And so that naturally is going to bifurcate development. We are becoming, in the West, more closed. There's basically less and less publication going on. And, you know, they're already very closed. Even though they actually publish a lot of good papers, they're very closed. So these chips are going to end up being the kind of main commodity that we negotiate over.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: And do you think the same in the area of synthetic biology?

MUSTAFA SULEYMAN: Well, hopefully with synthetic biology there are more immediate sort of incentives to drive cooperation. So, you know, it is going to be possible to synthesize with very few resources these new engineered pathogens. And that isn't in anybody's interest, their or ours. So there's going to be an incentive to cooperate on how we prevent those capabilities proliferating.

Number one, in knowledge and know-how on the web. And number two, which is basically going to be a global surveillance of the web to prevent this kind of thing. Just as we do already with preventing the spread of child sexual exploitation material on the web. And, you know, it's also going to be with restricting the export of these desktop synthesizers. I mean at the moment you can buy a DNA synthesizer, which enables you to print new compounds for, like $20,000. That's obviously going to come down in order of magnitude over the next five years. And, you know, this idea of being able to sort of buy the code for some new biological substance and then just print it in your garage is where we're going on the current trajectory.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: Could you elaborate on that because I'm not sure, there's been so much interest in AI, but the fact that DNA is really, it's information, and the fact that you can, you can now alter it, you can modify, you can adjust, you can create new information, and, as you say, you can synthesize, is really transformative in terms of everything that's happened before.

MUSTAFA SULEYMAN: Yes, I mean CRISPR is a moment in history that doesn't quite get enough attention. I mean really it's as significant as the arrival of these general-purpose language models. Maybe even more.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: Maybe even more, yes.

MUSTAFA SULEYMAN: Maybe more, maybe more. I think it just, it takes a little longer for people to see, okay, well, what can be done now that we can edit these snippets? But even if you set aside editing, simply printing known compounds and being able to do that extremely cheaply onsite, locally, is in itself a kind of, similar to the image model that I describe, a compression of power. It's sort of taking all the knowledge that we've collected over the decades, you know, knowledge in the kind of information space sense about what DNA is and how it can be used to produce new compounds and compressing that into a file that can be emailed around and then using that to actually print new substances that you kind of, you know, just on a whim. I mean that's a kind of type of power that we've just not grappled with in terms of its consequences so far.

So I think it's as significant. And I think it's an area where we want to, where we have good reason to cooperate with China, where we can provide, for example, some of our safety techniques. I mean why would we not share the tricks that we've invented for preventing these large language models from generating coaching for how to manufacture anthrax. I mean that seems to me like an area of cooperation that we can get behind.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: Well, the Chinese reacted quite firmly to their own scientists working with bioengineering or synthetic biology. And, in fact, I don't know what has happened to those scientists. They've disappeared.

MUSTAFA SULEYMAN: That's part of the problem. Yes, look, I mean obviously they have a much more authoritarian tendency and they want to shut down anything that looks like it could potentially be a threat. So, you know, you're right. This was three years ago. There was a moratorium by most of the western scientists because a scientist – I forget the name – had edited a human genome and given birth to...

CHAIR EMERITA MARIE-JOSÉE KRAVIS: …to a ___ . But is that the future?

MUSTAFA SULEYMAN: I think, yes.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: Designer genes.

MUSTAFA SULEYMAN: For sure. We have to accept that the default trajectory is proliferation. Because everything is getting cheaper and easier to use, the set of people who might have the intention to experiment like that, but previously didn't have the financial resources or the technical expertise, when you take those two things away, even if it's one in 100 million people on our planet who are sort of psychopathically motivated to engineer the next strain of Covid that is ten times more transmissible, that person will be found by these reducing costs and complexity curves.

And so we have to, you know, it's not doom-erism, as people often say, in Silicon Valley, where I'm from, to deeply engage with the potential of very dark outcomes. It's very rational. And it's actually critical that we start talking about it now, that we adopt a precautionary principle, and we take the possibility seriously. Because it is very possible. So there are plenty of interventions that we can make but they require action pretty soon.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: But we keep saying "we." But who is the "we"? Is it the government?

MUSTAFA SULEYMAN: Well, governments are, I think, faster than ever before. So I mean the White House issued a set of voluntary commitments to the top seven technology companies, Inflection included, it was one of the seven, to the largest developers of language models requiring them to proactively sign up to a set of restrictions about model uses, about the sharing of commercial IP, about how we keep our models safe. So if we make a safety breakthrough, we should share it with others. That we should publish reports describing our current capabilities, our underlying training data, etc. etc.

So I think, look, it's still a voluntary commitment, but it's quite a significant first step. And I think there's going to be an executive order out in a month or two enshrining those. And, you know, there's all kinds of regulatory efforts that are spawning. So I think the "we" is us as a species collectively. It's not just the model developers, but it actually involves everyday people paying attention to this question, you know, genuinely participating in the accountability process, giving feedback, saying what's acceptable and what we think is not.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: One of the recommendations in your book is that corporations spend 20% of their R&D expenditures on safety checks. Is that realistic given the competitive nature?

MUSTAFA SULEYMAN: I think it should also be audited as well, yes.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: So you think it should be compulsory, not voluntary?

MUSTAFA SULEYMAN: Yes, I do. I think it should be compulsory. It's completely rational. I mean, you know, we're inventing technologies with unprecedented power. I can't imagine anyone who wouldn't agree with that. And, you know, the reality is that, like we need innovation on the company side, on the academic side. We need movements and social organizing. And obviously we need innovation on the political side. I mean, for what it's worth, we've incorporated Inflection – my co-founder, Reid Hoffman, and I, and our third co-founder, Karen Simonyan – have incorporated, the company as a public benefit corporation.

And so this is a new type of corporation. It's not a B Corp, similar to a B Corp, which tries to reconcile this singular focus on returns to the shareholder, which is still our number one priority, with a fiduciary responsibility to satisfy the mission of the company, which is obviously public, which is safe and ethical HAI. And, you know, to factor in the consequences of our activities on people widely affected by what we do. So not just our customers, not just our supply chain, but anybody who might be affected. So you don't have to use our product for us to have a fiduciary responsibility to consider those potential future effects on those people.

Now, that doesn't solve all the problems. It doesn't mean that we will or we'll make the right decisions, but I think it's just a step in the right direction given the kind of magnitude of the things that we're inventing. And to me, it seems like a sensible operating philosophy for all companies to think about that. You know, you mentioned earlier...

CHAIR EMERITA MARIE-JOSÉE KRAVIS: Social responsibility.

MUSTAFA SULEYMAN: The social responsibility, yes.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: So let's talk about Inflection AI because that's my dream, is to have a Pi that certainly supports everything I do. And I'm sure everyone in this room would love to have a Pi. Describe Inflection AI, the product, and where you want to take the company.

MUSTAFA SULEYMAN: Yes, so I think there are going to be lots of different AIs. That's the first mental model to level-set on. There are going to be thousands and thousands of different AIs that have different types of expertise and are useful in different contexts. So, you know, just as we've had, like the arrival of the internet-produced websites, and then the arrival of mobile-produced apps, these are different representations for exchanging information, previously manifested in pixels.

Now, the next wave is going to be about interaction. And that means that your digital world is going to be mediated through a conversation. Rather than sort of just clicking on buttons and typing things into a keyboard, you're just going to say to your AI what you want and what you need and what you think and what you're missing. Just as you would say that to an assistant or a chief of staff or project manager, someone that you might work with. And that AI, your personal AI, is going to interact with the many other AIs that there'll be in the world.

So every brand, just like it has an app for a website, is going to have an AI. It's going to be represented – its values, its ideas – is going to be represented by its own AI. Just like every government has a website or a social media account, it's going to be represented by an AI. Every influencer or celebrity or musician, you know, every academic, every nonprofit group is going to represent its values through its own AI.

And in that world, you sort of don't – as a consumer and an individual person – don't want to be on the receiving end of these very powerful persuasive AIs talking at you. You want to have an equally powerful AI, an interlocutor, a translator, who can help you make sense of the world around you but is really aligned to your interests. And this is crucial. One of the lessons of the last wave of social media is that you are the product. You have become the product. And I think people forgot how fundamental that is. I mean, you know, if you're not paying for it, somebody else is paying for it. And that person or organization will be broadly aligned with your incentives, but they are never going to be 100% aligned with your incentives. That's a fact.

And I think we have lost sense of that basic, simple fiduciary connection between you and the service. In this age, if you don't have a direct one-to-one alignment between you and your representative, your digital representative that is going to play a very intimate part in your life, whether you like it or not, it is going to know a huge amount about you and be able to act on your behalf. Ultimately it will enter into commercial agreements on your behalf and make decisions for you. You know, it seems to me very clear that a personal AI is a critical piece of where this technical evolves, and that you should pay for that.

That's the crucial distinction in terms of the business model and lessons from the past. You know, you wouldn't allow someone else to pay your lawyer, right? Or your accountant, right? There's a fiduciary connection that you want there for a sensitive role. And I think that's going to be a very important distinction. That's what a personal AI is and that's what we're developing with Pi, so Pi stands for Personal Intelligence and also 3.142, etc. And we think everybody is going to have their own Pi, their own digital representative.

You know, the interesting thing about AI is that they are going to represent many, many different roles, which were previously independent. So your AI is going to both be a teacher to you, an amazing source of knowledge and information personalized to you. But it's also going to be a coach and a confidante giving you feedback and emotional support and advice. But it's also going to be a project manager. It's going to be prioritizing and planning and organizing.

And I sort of think of it as these three qualities: IQ, the models are getting more and more factually accurate, less biased and more smart; EQ, the models are getting emotionally intelligent and as we're starting to see now they really do provide very fluid sensitive, supportive, interactive exchanges, ask great questions. They can remember what you've said over time, paraphrase what you've said, be very kind and respectful. And then the third component is AQ, the actions quotient, like the models are going to start to be able to use tools, make phone calls, send emails, use APIs. And that's the kind of next wave that's coming in the next two years.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: But Pi hasn't actually be deployed.

MUSTAFA SULEYMAN: Pi can't do any actions yet. No. Pi can't do any actions. At the moment, Pi just has IQ and EQ. It's an app that you can use on iOS and on Android and ten different services. You can actually access it on Instagram, on WhatsApp? The concept is that an AI should be with you wherever you are. You should expect to have access to your personal intelligence whatever you are doing, wherever you are.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: I see Barbara. I think we've run out of time.

Could I just add one quick, quick question, just to close the loop? Because you began by saying, by talking about your interest in climate, and to what extent do you believe or do you aspire to see AI help in resolving some of these big climate issues?

MUSTAFA SULEYMAN: Yes. I mean I think there's a number of areas. The first is that this is going to be an amazing synthesizer of information. And that is one of the biggest challenges that we've had in the COP negotiations for 30 years, is distilling knowledge down into consumable nuggets that we can take action on, and I think that's very clear.

As a research assistant, it's very clear that these models are going to be – and I think already are – very good at coming up with hypotheses, finding evidence to support those, re-framing information in different ways. You can adjust the extent to which the model is creative or very specific with recalling exact pieces of information. That's a very useful research aid as we're inventing new things. So I think, if you think about it like that, as a scientific aid, I think it's going to have a huge impact on what we're able to create in the next few years.

CHAIR EMERITA MARIE-JOSÉE KRAVIS: Let's hope. Thank you, Mustafa, for sharing your ideas.

PRESIDENT BARBARA VAN ALLEN: Yes, thank you both. I think it's fair to say that's one of the most fascinating conversations we've ever had in the Club. So thank you for that.

Just real quick, because I let the schedule run a little bit, next Tuesday, October 17th, we have a breakfast with Pat Gelsinger, the CEO of Intel. He'll be in a conversation with John Williams, our Chair. On October 19th, as many of you know, we will be hosting Jay Powell, the Chair of the Federal Reserve, so tables are still available for those interested. On October 25th, we have a webinar with the President and CEO of Northwell, Health, Michael Dowling. And I'm not going to go through the whole schedule for November, but let me just mention December 7th, we will be having our closing dinner with Bill Gates, who will also receive the Peter G. Peterson Leadership Award that the Club gives annually. Please be sure you try to view our new podcast, The Forum, hosted by Becky Quick. It's available on all your favorite platforms.

And, of course, we always close by thanking the members of the Centennial Society that are here or online joining us today as their financial contributions help to make our programming possible. So thank you everyone online. And everyone in the room, have a great day. Thank you.