THE BROOKINGS INSTITUTION

FRONTIER AI REGULATION: PREPARING FOR THE FUTURE BEYOND CHATGPT

WEBINAR

Thursday, September 14, 2023

Washington, D.C.

UNCORRECTED TRANSCRIPT

OPENING REMARKS:

BEN HARRIS
Vice President and Director, Economic Studies
The Brookings Institution

INTRODUCTION:

SANJAY PATNAIK
Director, Center on Regulation and Markets
The Brookings Institution

FIRESIDE CHAT:

THE HON. TED LIEU (D-CALIF.)
Member, U.S. House of Representatives

MODERATOR:

ANTON KORINEK
Nonresident Fellow, Economic Studies
The Brookings Institution

PANEL:

GILLIAN HADFIELD
Chair and Director
Schwartz Reisman Institute for Technology and Society

DAN HENDRYCKS
Executive Director
Center for AI Safety

ADAM THIERER
Resident Senior Fellow, Technology and Innovation
R Street

* * * * *

HARRIS: Oh, good afternoon, everyone, and welcome virtually to Brookings. My name is Ben Harris and I'm the vice president and director of the Economic Studies program here at the Brookings Institution. I'm very pleased that you all have chosen to join us for this extremely important event. We're here today to discuss the regulation of frontier AI systems to prepare for the future beyond ChatGPT. We are honored to have several high-level experts with us today who are particularly well-equipped to provide their insights on the topic. For a keynote address, we are very pleased to welcome Congressman Ted Lieu, representing California's 36th District. As for our panelists, we are also very happy to host Gillian Hadfield, chair and director of the Schwartz Reisman Institute for Technology and Society; Dan Hendrycks, executive director, Center for AI Safety. Adam Thierer, resident senior fellow, technology and innovation, R Street Institute; and our moderator will be Anton Korinek, resident fellow, Economic Studies, for the Center on Regulation and Markets. My thanks to each of you for taking part in this conversation.

As we all know, the current landscape is complex. We're seeing groundbreaking technological advances with systems like ChatGPT. These innovations signal enormous potential for society. The promise of greater productivity, more dynamic and fulfilling work, and technological progress that improve lives. However, as we look ahead at the rapid evolution of AI, we must also recognize the significant risks and challenges these systems present. As AI advances and acquires more humanlike intelligence across diverse domains, we face heightened dangers of misuse and unintended harm. For example, misuse of frontier AI could enable bad actors to orchestrate criminal activity or violence on an unprecedented scale. Even with positive intent, accidents stemming from the complexity of advanced systems could spiral out of control and endanger public safety. We cannot afford to ignore or downplay these risks.

That is why at today's event we will discuss concrete ways to evaluate the future of AI governance. We will address how policymakers, companies, researchers, and the public can work together to maximize AI's benefits while developing prudent safeguards. We will also explore what policies and practices will allow innovation to thrive while minimizing negative consequences. By starting this discussion now, we hope to get ahead of the risks and create a trajectory for AI that is ethical, equitable, and aligned with human values. There are no easy answers, but through collaboration and open exchange of ideas, we can pave a responsible path forward. I very much look forward to the discussion today and we'll now turn the proceedings over to Sanjay, the director of the Center on Regulation and Markets in Economic Studies. Over to you, Sanjay.

PATNAIK: Thank you so much, Ben. And welcome, everyone. I'm the director of the Center of Regulation of Markets here at Brookings, and this event is part of our workstream on AI and emerging technologies that we have been really pushing forward over the last three years. As Ben mentioned, the regulation of AI and new emerging technologies, as well as the economic impacts, are becoming more and more important as this technology becomes more mature, and as more and more markets get affected by it. So, it's really wonderful to have this great panel of experts here today. And Congressman Lieu is currently in Congress finishing up some votes, but he will join us very shortly. I'll just briefly run through his bio for a few minutes so that you know what his background is before handing it over to Anton, who will moderate the fireside chat with him as well as the panel afterward.

So, Congressman Lieu represents California's 36th Congressional District in the U.S. House of Representatives. He is currently in his fifth term in Congress and currently sits on the House Judiciary; Foreign Affairs; and Science, Space, and Technology committees. He was elected by his colleagues to serve as vice chair of the Democratic Caucus, making him the highest-ranking Asian-American to have ever served in House leadership. Ted is a veteran having served in active duty and then in the Reserve for the Air Force. He retired from the Reserve in 2021 with the rank of colonel. Congressman Lieu is also one of the few computer science majors currently serving in Congress. And that means his insight on technology and innovation matters including artificial intelligence, really have had significant impact.

Anton, I want to hand it over to you so that you can start off with some starting words before the congressman joins us after his vote. And thank you again for moderating this wonderful high-profile panel of experts and especially talking about this really important topic that we all care about. Thank you.

KORINEK: Thank you very much, Sanjay and Ben, for the introduction and for organizing this event on what is perhaps one of the most important topics of our lifetime. How to control the evermore advanced systems that are developing at such a rapid pace? I think there is a recognition all around the world in governments, corporations, think tanks, universities, really across all of society that there are groundbreaking advances in AI going on. These advances have raised a lot of challenges for existing regulations. And there are many conversations about how to best reap the benefits. For me, I like greater productivity, more enjoyable work, while also minimizing the downsides and potential harms like lack of robustness, bias, deteriorating working conditions, and so on. However, what our conversation today is about is frontier AI systems. Not the myriad of AI systems that we already have all around us, but the systems that are at a very cutting edge and in particular even more powerful systems that are currently in development that have not yet been created, but that we can see on the horizon because the pace of advances in this sector is so fast.

So, why would we care about what is on the horizon in the midst of all the challenges that we are facing with today's technologies? Well, to me, an important answer is that intelligence confers power. We humans, we are the most powerful species on planet Earth. Not because we are the strongest or because we are the biggest, or because we are the fastest, but because we are the most intelligent. And as AI is advancing and acquiring more and more humanlike and super humanlike intellectual capabilities across a growing range of domains, they are also becoming more powerful. So, it means two things, two types of risks: the risk of misuse and the risk of accident. The first one, misuse risk, captures that if someone wants to employ these frontier AI systems for nefarious goals, for criminal activity, violence, and so on, they may be able to create havoc on an unprecedented scale because of the power of these systems. The second type of risk, accident risk, reflects that if these powerful systems function in ways that we did not intend, the consequences will be ever greater. They will be commensurate with how powerful they are, with how important of the role to play in our economy and in our society. And ultimately both the misuse and the accident risk may pose great dangers to our public safety and well-being.

There is a significant amount of uncertainty about the risks posed by frontier AI systems, and even leading AI researchers are unsure about when they will start to materialize, how soon they are on the horizon. But given the large number of experts who have expressed serious concerns, and Dan, who is on our panel today, has collected a large number of signatures from experts on this topic who all agree that we should be concerned about the potential risks from frontier AI systems, about even the scope for potentially existential risks. So given this significant degree of uncertainty, it is important to be prepared. And it is important to have a public debate on what policy measures may be most useful to address these risks. At the same time, I should emphasize that it is also important to keep in mind all the potential that advances in AI offer for us. They offer the potential for prosperity that we have never experienced before. And so, we want to make sure that we balance any regulations that we may be tempted to impose to forestall safety risks with the potential that we do not want to curtail.

So, I just received a message that Congressman Lieu is available to start our keynote fireside chat. And, Congressman, it's an honor to welcome you back to Brookings. And thank you for joining us for this fireside chat on frontier AI regulation. There has been a lot of debate about AI regulation in Congress recently, but you've been a trailblazer. You were really the first person in Congress to call attention to the disruption that may lie ahead. And you have, for example, written an op-ed in January of this year already in the nation's most prestigious newspaper that was entitled, "I'm a Congressman Who Codes. AI Freaks Me Out." Can you lay out your main concerns about frontier AI systems for us in a little bit more texture?

LIEU: Thank you so much, Anton, for your question and for moderating, and to Brookings for having me on. I apologize, I was a little behind. They called votes on the House floor at 1:30, so I'm actually calling in from the cloakroom of the House floor. To answer your question, I'm gonna tell you how I view AI from the perspective of a lawmaker. I think the best analogy I have is to think of two bodies of water. You have a large ocean of AI and then a small lake of AI. So, in this large ocean is all the AI we don't care about. So, if there-- AI in your smart toaster does a better job and has a preference for bagels over rye toast, we don't care.

Now, in this small lake of AI is the AI we might want to care about, and to me there's three buckets of why we want to do that. The first is AI that can destroy the world. So, in the Department of Defense, there are weapons known as autonomous weapons that can launch automatically. I've introduced bipartisan legislation that basically says no matter how amazing AI ever gets, we're never going to let it launch a nuclear weapon by itself, as we humans [inaudible]. We're currently working on legislation that will try and mitigate the risk of finding out from a large language model how to make the next virus that's gonna cause a pandemic. Because unfortunately with biological weapons, they are significantly easier to make than, for example, a nuclear weapon where you have to enrich uranium. You can in fact make a virus and there are companies that do that. They will literally send you a vial of the 1980 virus that caused that flu. And now they shouldn't send to you. They should do a background check and so on, but the capability is there and it's much easier. And so, we want to make sure that generative AI is not going to allow terrorist groups and nonstate actors to know how to make viruses much easier.

Second bucket is the AI that's not going to destroy the world but can kill you individually. So, when your cell phone malfunctions, it's not going 50 miles per hour. But if a, an AI-automated vehicle malfunctions, it can kill you. It has killed people. It's going to kill more people. It's a lot of AI in moving objects: planes, trains, automobiles. I think there should be more regulators at the state and federal level more attuned to unique aspects of AI.

And last bucket is the hardest, which is any AI algorithm that does have widespread societal harm, but that doesn't kill you. So, there is biases, for example, that we do care about. Whether it's in algorithms that do credit risks, or that deal with facial recognition, or hiring, and so on. And that's how I would view it. But there's both benefits and costs to overregulating. So, we have be careful how we tread.

KORINEK: Thank you. Congressman, you have introduced a bill, a bipartisan, bicameral bill, to create a national commission on artificial intelligence in June. Now, we currently have a divided Congress. So, the bipartisan and bicameral nature of your bill is very important. Can you tell us a bit about the objective of this bill? And can you speak more broadly to the question of why you see concerns about frontier AI as a bipartisan issue?

LIEU: Sure. This bill would create a national blue ribbon AI commission of experts to make recommendations to Congress as to what kinds of AI we should regulate and how we might want to go about doing so. There's precedents for this. There was AI commission on the military side that made recommendations, a number of which were adopted. So, this is on the civilian side.

And there are some reasons why I think this would make sense. First of all, if you were to say, "Hey, let's regulate AI next week," I don't even know if we would be able to define what that is. So, I think we need to get more understanding from experts and also have some time pass so that we actually see, do all 67 harms predict from AI in fact happen. Maybe none of them do. Or maybe it's three harms and one really huge harm we haven't even thought of.

And then another reason is, it's going to be fully transparent. So, I realized pretty quickly it's not really helpful to the American people if I talk to 37 AI experts or I talk to Silicon Valley titans because you have no idea what they told me. This commission would be public, transparent, you'll know who they talk to, what information they relied upon, and how they got to their conclusions. So. those are the reasons why I think we should have this commission to look at frontier AI.

KORINEK: Now, I think one of the proposals that you advanced is that the commission should basically do the preparatory work to establish an agency for AI oversight. And I think that's a very important step because the complexity of establishing a new agency is vast in the complex world that we live today. So, I have in fact been arguing that we need such an agency for over two years. But many counter that we don't need an AI agency, we already have regulatory agencies that can, for example, to pick up on your previous example of biotechnology, that can regulate biotechnology. We have agencies that can regulate self-driving cars. What do you view as the benefits of ultimately establishing a dedicated AI agency and why should we pursue that?

LIEU: Well, let me first start with the broader question of why a approach based on agency or agencies and regulators would be better than passing specific laws on every application. So, I introduced a bill on facial recognition because, with the current technology, it is less accurate for

people with darker skin. And my view is if you don't put guardrails on that and you deploy that at law enforcement agencies nationwide, it is a huge equal protection violation because minorities will be misidentified at higher rates. It also took me over two years working with stakeholders to put out a bill that makes sense.

So, it's very clear to me that Congress does not have the bandwidth or the capacity to regulate AI through individual laws and every possible harmful application. I mean, we're just trying to stop stupid stuff from happening right now, like not have the government shut down, right? So, that's where we are. Now, what the advantage is with regulators is you have experts thinking about these issues, working on these issues, every day of the week in a way that Congress is unable to. And if regulators make a mistake, you don't need another act of Congress to fix it. Regulators can correct themselves. If we pass a law and make mistake, we need an act of Congress to fix that. So, I think regulators are also much more flexible. Now in terms of whether it is sort of one overarching agency like the FDA, for example, and how the FDA regulates drugs, or if it's empowering, you know, eight different agencies with more power over their specific sectors in terms of AI, I'm agnostic on that issue. That's why I would like to see the input of national [inaudible] experts to provide reasons for why one approach might be better than another one.

KORINEK: I see. So, you are proposing that the expert commission should investigate the issue and then make proposals based on that?

LIEU: Yes, it’s very clear to me that it'll be impossible for Congress to keep trying to pass individual laws on every single possible harmful AI application.

KORINEK: Great, thank you. Now, in the spirit of discussing policy proposals that may garner bipartisan support it seems like one of the fundamental problems that we are facing at this juncture is that government has very little visibility on what's going on at the very cutting edge of AI. So, a number of AI companies have recently agreed with the current administration that they will inform them on what's going on with basically their most cutting-edge AI systems. But this is all on a completely voluntary basis.

So, I wanted to ask you, what do you think about proposals to establish simply monitoring regimes to make companies register their most cutting-edge AI systems that they are developing with the government to keep track of who holds the most advanced AI chips that are used to train the most cutting-edge models so that we establish a basic level of visibility around these systems for government.

LIEU: I will be open to that, although I am not sure how much it is needed in this sense? Very few companies and countries can do these large language models. So, ChatGPT basically had 25,000 super expensive Nvidia chips in order to run-- train the model and to run it. And the power they use can be seen from space. So, we basically know what companies are doing this. It's sort of hard to do this without people who're knowing you're doing it. And again, very few companies and countries right now can scale in this manner. So, I'm not sure how much this would tell us other-- any more than what we already know. I'm open to it but I'm not sure that it would be a huge utility, but I could give it some some more thought.

KORINEK: Great, thank you. One, one more question. Four months ago in May, we both signed the statement on AI risk that one of our experts in the panel, that will follow this fireside chat, prepared. I'll read the statement: "Mitigating the risk of extinction from AI should be a global priority alongside other societal scale risks such as pandemics and nuclear wars." Now, you personally, you are an expert on AI, and you understand both the opportunities and the risks arising from it. But for non-experts, it's sometimes difficult and confusing to appreciate the opportunities and balance them with the risks. How do you explain the risks and the opportunities of frontier AI to your constituents?

LIEU: So, as a recovering computer science major, I am enthralled with AI. I think AI already has helped society. It will continue to move society forward. And in general, it's going to provide a lot of benefits, particularly in the medical field. So, for example, it used to take a human being five years studying a Ph.D. to tell you how to fold one human protein. Now, AI has folded every single human protein known humankind and given that to medical researchers. So, I think we're gonna see some tremendous healthcare advances, as well as in other fields.

At the same time, there are risks. And the statement you mentioned identify two, right? Nuclear weapons and pandemics. So, AI would increase the risk of nuclear war if you gave AI the ability to launch nuclear weapons by itself. And I think that is a problem and I think we need to make sure that no matter how captivated we are by AI, or amazed by it, or so on, we can never have that be the case. It always has to be human or loop. And in terms of pandemics, there was a experiment run by MIT professor, Dr. Esvelt. He basically told his students to go on generative AI Applications and try to figure out can they design a harmful virus? And after about an hour, they, they basically came up with a way to do it, which is quite frightening. And so, I think we have to consider whether we should have government agencies be able to go in and tell these AI companies, “Look, you can't do this" or "You got to have better guardrails on how to prevent that from happening."

I think that that also raises the issue of open-source versus closed-source. So, ChatGPT is able to basically, right, put in guardrails and when it finds someone that's tricked it, then it can do countermeasures and there's probably been this ever-constant battle every day on, you know, overcoming guardrails and putting them back in. But at least ChatGPT can do that. If you take the approach, for example, of some companies like Meta, that put out open-source generative AI like Llama 2, where the guardrails can be removed much easier, there's a question of whether that should be allowed or not. So, I have to ask that question and I'm trying to get input on it. I asked Secretary Energy Jennifer Granholm too, that question because her office has jurisdiction over nuclear proliferation. And I'm still open to what the answer is, but I think it's a legitimate question we are to think about whether we should have very large language models with no guardrails because that's essentially where open source is going to lead to.

KORINEK: Yes, indeed. And if I may add, one of the things that this open-source model Llama has already enabled researchers to do is to break down all the guardrails of the other language models that have much better guardrails. So, I have one last question for you as we conclude our conversation. We all hear from a growing number of people who learn about the risks of advanced AI, and I wanted to ask you: what can our listeners do to contribute to meeting the challenge arising from

that? Can you perhaps propose one small but tangible action for everyone in this call who wants to have a positive impact to mitigate the risks of the AI?

LIEU: Well, you know, this is just purely personal but, you know, lobby your Senators, Representatives, to support my bill. But setting that aside, I think it's important for people to continue to raise this issue because everyone watching, I'm sure, is aware of generative AI, has a lot of knowledge about this. But most people are just trying to make ends meet and live their busy lives, and they may have no idea what generative AI is. They may never see ChatGPT. They're trying to just pay their bills. And so, just highlighting the issues for people to see and understand, I think is extremely important. I think these programs run by Brookings is important. I think then getting this out for people to watch on social media or other ways to see it would be important. But just highlighting the issue, because I want more Americans to understand what generative AI is, what it isn't, what it can do, and both its benefits and its harms.

KORINEK: Thank you, Congressman. Thank you for this uplifting message. And I think that it's important for everyone to hear that our representatives are listening to our concerns. And if we have these concerns, we can always raise them with them.

LIEU: Great. Thank you very much.

KORINEK: Thanks again.

LIEU: Have a good day.

KORINEK: Have a good day. And we will now continue with the rest of our program. So, I interrupted my remarks before because the Congressman could speak right in between two votes for which he had to be present on the House floor. But I was just about to speak about the necessity of balancing the potential upsides of AI with the downsides that call for regulation. And two of the important downsides that I think we will be discussing in more depth for the rest of our hour together, are that regulations may create entry barriers that could entrench existing players in the industry, and it could make our industries less dynamic. And they also run the risk of undermining the competitive position of the U.S. So, whenever we advocate regulation, we have to balance that with the potential risks of that. But of course, both the risks of no action on the regulatory side and the risks of excessive action on the regulatory side are tremendous.

So, I think what I want to conclude with before we start with the panel conversation is that governments around the world have now started to pay attention to this question of what should we do with frontier AI systems. The Congressman has just spoken to the necessity of thinking about what advanced language models can do and whether we should even allow them to be open-sourced, for example. And getting this balance of regulation right is one of the most important challenges that we're facing in this space.

So, I want to now proceed by asking the panel on stage, and I want to start, maybe we will just go in alphabetical order with Gillian first. And I want to start with a few questions to Gillian, and then Dan, and then Adam. Gillian, you have been active in this debate for a very long time. And recently, you have contributed to several proposals for the regulation of really the most cutting-edge AI systems, the ones that I call frontier AI systems, ranging from more modest proposals like registration and monitoring requirements, to more complex proposals like what you and I proposed in

a recent paper that was entitled, "Frontier AI Regulation." Can you give us a quick overview over these proposals over how you see the timeline for each of them, and what the benefits and disadvantages of each are?

HADFIELD: Yeah, thanks. Thanks, Anton. So, so I think, you know, I want to pick up on a point that the senator made is about-- that we we can think through what possible risks, and I think there are a lot of things for us to be thinking through. We've been thinking for many years about risks of discrimination, privacy violations, and so on. And as we start to think about frontier systems, thinking through misuse, as you mentioned, accident to Dan is drawing attention to sort of the, you know, the broader extinction risks. I think about system collapse risks. Those are the kinds of things that I spend a lot more time on. And I think that point about the fact that we we can imagine the scope, but we don't really know yet. And I think it will take time for us to understand because systems are evolving so fast and then they get out into the world, and they get used in ways. That's what a decentralized market economy does for you, is that there's things that happen you don't know about.

And so, so just to sort of think about, you know, the way I've been thinking about this, and I'm going to start with our proposal, proposal that I've made with Tino Cuéllar and Tim O'Reilly, for creating a registration requirement for our largest models. And to see that is actually a seed crystal for creating the kind of agency, for example, that you you've been talking about, and evolve to those-- I think we are going to need licensing requirements. We're going to need standards. We're going to need enforcement. We're going to need liabilities. But I think we are at a stage where we need to build the infrastructure for that.

So, let me just sort of quickly say that the, the registration requirement is not just an information gathering, right? It's, it's actually a requirement that in order to sell or buy, use the services of our our largest models, the frontier models, the model has to be registered with a national agency. Then I think that's the seed crystal for that agency. And registration requires disclosure to that regulator. Not just, "Hey, we're building this," but also what training, data, size, what we know about its capabilities, what kinds of tests we've done, and so on. And I think that's a level of visibility we've lost in this latest round. I mean, we know these things about GPT three, for example. We don't know them about the biggest models and what's coming next. And I think that that's again a requirement that it be, right? It's like our our requirement of corporations. In order to do business in the state, you must register your corporation. That means you've got an address; you've got people associated with it. And it might be like, you have a board of directors who've got corporate law that that governs that.

So, that's kind of a-- I see that as something we can do quite quickly, and I think it should be our first step. I'd love the idea that we recruit sort of top scientists and regulators to that, and regulatory experts to that, to that exercise of then figuring out, okay, what are the licensing requirements. So, in the paper that you mentioned on frontier AI regulation, we talk about, okay, there are standards, there's tests you should have to do. There's probably requirements about data. There are requirements about, we know there's a human labeling component that's going into the latest models. Where are those labels coming from? What are the instructions? The safety requirements about, about cybersecurity? Those are the kinds of things that we need to be evolving. And I think we

need that high-level visibility to to start with. I think that's where we will go. But with a registration regime, you immediately get a lever. You can say things.

So, again, Senator Lieu's talking about, you know, you don't want these models in the hands of certain users. Well, with the registration requirement, you can actually say, you know, you can't you can't maintain your registration if you're making it available to a particular kind of user or a particular state. So, I think, I think you actually get some regulatory levers immediately from just a registration requirement as you start to figure out where we... I think you can't regulate what you can't see. I do-- I don't think we know enough. When you put the question to Senator Lieu of of well, you know, "What about registration?" Well, we already know. And I would say, "Well, I don't think we really do." And it's, and the open-source question makes that its-- everything is evolving. We know right now. And I think we need a formal structure that gives us all that information.

KORINEK: Now, Gillian, as a follow-up, we spoke about the registration of models, but in our paper, we were also talking a bit about the hardware side of compute coverage. How do you see the registration of models versus keeping track of who owns the chip you use to compliment [inaudible]?

HADFIELD: Yeah. So, so I, you know, we've got that in the paper, and I'm, I'm working on another paper now to try and develop some of these ideas of of compute regulation. So, so I think, for example, you can imagine a comparable kind of registration requirement. And it could be that, that you now require your data centers or large providers of compute to register. And maybe there has to be disclosure of of use beyond a certain level. You know, we can, we can think about, I mean, there's, you know, there's tracking, there is the capacity. You know, we could have a regime that said, "You know, we're going to just keep track of where all those, particularly the chips you need for the most powerful systems, where they are, who's got them, and and who's building with them," as another way of gaining that-- one, visibility. But two, also, again, if that's your-- you know, I'm I'm a lawyer and an economist. I think about the design of markets and the design of legal infrastructure. You need to have those pieces in place to be able to say, "Oh, we've just figured out that we want to restrict the capacity of this criminal entity from from building a model or using a model." But building the model means that you, that compute gives you a lever to potentially say instead of saying, "Hey, you're not allowed to build that," you'd say, "Oh, and you're not allowed to access the compute you need for that."

KORINEK: Yeah. Thank you, Gillian. I have two questions for Dan next. Dan, we have already spoken about the, by now, famous statement on existential risk that you put forward at the Center for AI Safety. I wanted to ask you, what motivated you to initiate this kind of public statement? And did you expect the impacts that it had?

HENDRYCKS: Yeah. So, thank you for having me. A lot of the reason for putting together the statement was because I noticed that a lot of academics were concerned about this issue, but they were somewhat afraid to speak up. So, because of the changes with ChatGPT, a lot of them had substantially shorter timelines for when we're getting to very advanced AI systems. So, many people were changing their minds. But the public deserved to know that this was actually of this larger development. So, that's why we put it together. And we ended up finding that there were many signatories who were quite unexpected. So, there are many professors who were concerned about, or

are concerned about, even extinction risks from these systems that are outside of our network. So, it became a lot of, a lot of people unexpectedly, and as we've seen like yesterday, the president of the European Commission read the statement. So, it has been, it has been surprising the impact and it's very fortunate that we can start to have a discussion about these these risks because the time window could be fairly short if AI technology keeps progressing so quickly. We need to get a handle on it. We need to figure out shared standards and coordinate around this problem. So, it's been a very positive development, but there are many risks that we'll have to address now that we've established urgency and importance behind the risk. Now, we need to talk about what are some potential solutions and negotiate and find the appropriate balance.

KORINEK: Indeed. Yeah, in fact, I think looking through the virtual room, the majority of participants of this call have all signed the statements. Now, I wanted to ask you also, can you perhaps lay out how you see the risks of the most advanced frontier AI systems and how can you best convey that to non-experts?

HENDRYCKS: Well, so one, one thing I'll note is we have a paper on this that I'll link to potentially if people can see the Slack comments, but it's called "An Overview of Catastrophic AI Risks," where we try and lay out the variety of them. So, the congressman had mentioned using it for bioweapons, is one such issue. Or if we keep automating more and more to weaponized AI systems, including command and control, and communications, then we're in a substantial-- we put ourselves at some substantial risk. I think of it as you mentioned, you mentioned intentional risks, or malicious use, and accidental, the sort of breakdown I'll tend to go with is yes, malicious use. There's risks of of of accidents too: people, organizations, you know, moving fast and breaking things accidentally, things like that. Not having robust processes for important decisions. Just as nuclear power plants melt down and rocket ships explode, we could have similar types of issues with advanced AI technologies from accident risks. There's some other ones too, which would be these structural risks and risks that are inherently from the AI systems, which are maybe, are maybe the main ones I'd like to speak about in addition to the ones that have been discussed so far today.

So, with structural risks, this is where we're having extreme competitive pressures and we keep, we keep having to race and move as quickly as possible, AI developers do, even though they know that the risk is potentially quite high, and they would all like to not have to cut corners on safety to stay competitive. Their sort of reasoning goes as follows: "that, well, if we try and take things more cautiously if we stop prioritizing profit over safety and instead invest substantially more in safety, that would give more unscrupulous actors a leg up. So, we can't do that. Therefore, we're going to compete too heavily." But we can see this with like open AI. Open AI partly started out as a, it started out as a nonprofit and beneficial for humanity. And while they've, I think, have been possibly the most responsible major AI organization, they still had to, you know, turn into a-- cap prof a bit to raise capital, and to prioritizing, you know, being on the forefront quite a bit. Anthropic wasn't happy, or many Open AI employees weren't happy with this, so left to create Anthropic, which then had a strong safety focus. But then because of the intense competition and the need to race on these issues, then they ended up behaving fairly similarly as well.

So, I think this is what drives AI development stronger than any other factor. And they all would like to chart a somewhat different course, a more responsible, prudent course. But it isn't possible because there aren't shared standards, there aren't regulation, there isn't a referee. And I think that's partly one reason why many of these organizations signed this statement because we're facing a classic collective action problem where, well, we would all like to do this, but it would cost me too much individually to do it, just like, you know, with nuclear weapons. With nuclear weapons, nobody wants extremely large arsenals, but it makes sense for them to stockpile many of them.

But this creates a classic security dilemma. So, I think we're in a similar problem. And the usual solution for these would be some type of coordination domestically, and later on internationally too. This is another reason why it's important that the U.S. focus on pushing for these sorts of standards because that's the only thing that will enable us to coordinate with with other actors on this front too. China seems to be pushing on this independently. So, hopefully, there would be some room for working together. At a later stage, these these structural concerns can look like us just basically drifting into a state of extreme dependence on AI systems. So, imagine the market being fueled by AI. We've got increasingly powerful systems. One day we're letting it draft some of our emails for us. Eventually, it gets good enough that it's basically an assistant for us. Then we replace some people with these AI assistants. Maybe it starts having good marketing ideas too. I mean, this, we're speaking years out. But across time, we end up outsourcing more and more to them and become more dependent on them. The economy moves more quickly too. Things are happening faster than we can even read about.

So, we need these AI systems. We end up-- the solution to many of our problems is to use more AI systems. If we try and resist this trend, or if some people try and resist this trend or organizations, they are getting out-competed. So, there's some selection for the ones that really lean into this, this rapid automation process. So, we have less and less oversight over these AI systems. Everything moves faster and this makes the situation substantially more unpredictable. People have just nominal influence over what these AI systems do. They're doing most of the effective decision-making because we can't make the decisions as intelligently ourselves. So, this is a situation in which we lose effective control. Nobody is to blame. This wasn't because AIs maliciously schemed or anything like that, but because we basically succumbed to a gradual collective action problem, the most efficient, most competitive systems end up winning out. And unfortunately, as AI systems become more and more competitive, they will naturally replace us in many ways.

So, I want to make sure that we can keep, keep control of that process. And that after we've given them, you know, the keys to the kingdom, that we're we're still in a safe place so that it doesn't automatically swerve on us. Like we'd be building sort of an autonomous economy. Like an autonomous vehicle, you wouldn't want it randomly swerving on you unexpectedly. So, that's a structural type of risk. No one's to blame, but it's a collective action problem fueled by AI race dynamics, which could get even worse with an AI arms race, if it comes to that.

And then the other risk would be that of rogue AI systems, where we have a loss of control because maybe they had the wrong goal put into them, or there is something accidental, that they got some wrong goal or some some sort of emergent goal happening. And we know that they have

emergent capabilities. And we see when they come together they can have emergent goals when you have lots of AI systems talking to each other, they have some emergent plan. So, if they get some emergent goals and if they start pursuing those, then we're in substantial trouble because if they're smarter than us and more powerful than us later on, and have some goals separate from us, it's a very distinct adversary. This isn't like trying to make an airplane safe or anything like that. We've got a quite, quite frankly, quite the adversary that's smarter than us. So, I think they'd have a-- it would be very dangerous if that would happen.

Now, I think that the-- in my mind, these these four categories of risk: the misuse risks, the accident risks, the the structural AI arms race risks, and the, the other risk source of rogue AIs. I would think that accident risks and rogue AIs are somewhat lower priority for me compared to the malicious use and the structural risks, at least in the in the next few years. But I think we should be trying to address all of them for a-- I know many people are, you know, looking into this AI stuff now and wanting, you know, some, a lot of material on this. So, if you're wanting to just get a broad overview of it, then I'd suggest looking at the overview, the paper entitled, "An Overview of Catastrophic AI Risks," where we try and just get all the main important points across. And we suggest different policies, but we try and help people get a more complete picture of the various risks that this could pose.

KORINEK: And thank you, Dan, for this very comprehensive overview. And as an economist, I will add to the structural risks, the risk of really massive job disruption, not at this point, but as these systems can power the economy more and more without human input, as you described. And I can see that to be a really dramatic risk factor for our society as well. I'll turn to Adam now. Now, Adam, you have also written extensively about the risks of AI systems and the proposals that you have advanced have taken a slightly different angle, if I characterized it right. You have advocated self-regulation as one of the solutions to what we are seeing. Can you tell us a little bit more about first, what you are proposing, what the benefits of self-regulation are? And then in a follow-up, I will also ask you to compare it a little bit more with regulatory proposals that actually put the force of law behind what is required of companies.

THIERER: Well, thanks for having me, Anton. It's great to be here. So, let me just take a step back here and actually talk about sort of my normative concerns and practical concerns about what's being proposed for frontier AI regulation, because I fear at this point in time it appears that AI policy is now threatening to devolve into an all-out war on computation and computing. We are looking at some of the most sweeping licensing regulatory regimes ever proposed. We are looking at national and global coordination or regulatory bodies for AI computation. We are looking at controlling the flow of chips, the movement of chips, the development of them, the data centers that run them. We are looking at regulations, if not sweeping bans, on open-source technologies. There's even talk of widespread surveillance of scientists and others engaged in AI-related R&D exercises across the globe. At the even more extreme, we see calls for nationalizing various types of computational systems. The person who's currently been appointed to run the AI safety effort for the U.K. government has basically proposed the idea of putting everything on a single, quote-unquote, AI island to control. Who knows who runs us, who knows where it's at. But somehow we're going to just

put the most high-powered systems in the world all together in one little mountain hideaway somewhere in, I don't know, maybe Zurich, maybe somewhere else. But that's that's just incredible to me. And I mean, we even have these this casual talk about like, "Well, what about bombing data centers?" People published in Time magazine saying such things.

If you ask me, it sounds like we're substituting one sort of very real existential risk for a very hypothetical existential risk. I mean, we know one thing. Global governmental controls and totalitarian kind of systems for our technology are a real existential risk, and we should avoid them at all cost. But basically, we're talking about the great recentralization of all information and communications technologies and resuscitation of the crypto wars that we fought in the late 1990s with widespread controls on computation. That's just an enormative concern. Practically speaking, all of this debate operates in the world is sort of like abstract, aspirational statements. You know, let's "pause" AI, whatever that means. I don't know what that means. Pause what, when, where, and how. Who's pausing? What if you can do something beyond the pause? Who's going to enforce you to stop it? These things are just hypothetical, like proposals that have no meaning. But let's talk about if they had meaning. Practically speaking, how are we going to get everybody else in the world to go along?

Just some quick data points here. Just just recently, a couple of weeks ago, it was announced that China is now up to-- on the top 500 supercomputing centers in the world, 227 of them, 45% of the world's top 500 supercomputing systems, are in China. The U.S. is at an all-time low at 118, now under 25%. Oh sure, China's looking into AI safety, and we might be able to believe that song, but I don't. And I don't think we're going to have any global international body in the U.S., U.K., or Europe, that they're going to come and sit at the table and be an honest player at. What about Russia? They just announced two weeks ago that they had developed one of the most powerful supercomputers they've ever, ever developed. Don't think Vladimir is going to come to the table and bargain on this, but how about something more simple? How about the the UAE, which has just developed and released the latest edition of Falcon? Falcon-180B, 180 standing for 180 billion parameters, which has now already dethroned Meta's Llama model, which is a 70 billion parameter model, that just two months ago was the most powerful open source model in the world. I suppose we could control Meta in the U.S. Are we going to be able to control, you know, a UAE-based, open-sourced,180 billion parameter model from the U.S., U.K., Europe? I guess we can try, but I just don't believe this is realistic.

I also don't understand exactly what our threshold or metrics are. There's some talks about, well, a certain level of, you know, parameters and and tokens, or maybe it's based on petaflops or something else. I don't see a concrete metric that's emerged that makes sense, and I don't see how it's enforceable internationally. Now, look, despite all of these things I've said, that I expressed obviously profound skepticism, concern about these proposals, I don't believe in anarchy. I believe in regulating risks. I believe there's already a lot of laws, policies, and systems to regulate those risks. I spent a lot of my time writing about how AI in the real world, in real-time, is being regulated right now to address existential risks. You want to talk about the existential risks that AI and ML related healthcare, medical devices? Well, that's my next paper. The FDA's all over it. They've been on it for many, many years under different names: mobile health, digital health. But now they're regulating it

very aggressively. You want to talk about driverless cars? NHTS is on top of that. You want to talk about FAA regulating drones? You want to talk about all the different alphabet soup? Four hundred and thirty-four federal agencies in the United States, 2.2 million employees working for them. You want to tell me that nobody's interested in AI? The FTC, the DOJ, and many, many other agencies are all over this. At the state level in the United States, I've lost count, but somewhere around 100 bills pending right now or have advanced. And then you have city and municipal regulations. And then you have all the international stuff. And some of this makes sense. A lot of it may be excessive, in my opinion.

But the bottom line is, let's not pretend we live in a state of anarchy. We don't. We are aggressively pursuing AI policy and regulation, but just in different names and in different contexts. And that's the way we should do it. AI regulation should be focused on outputs and outcomes, not inputs or systems designs. We should not be micromanaging from above, either in Washington or some far-off global capital, exactly how algorithmic and computational systems are designed. Yes, we should coordinate. Yes, we should have safety standards. It should be more than just best practices in some cases. I'm all right with certain types of registration and so on and so forth. We should have at least many lateral processes where we talk about how to deal with the most serious risk. I don't think multilateral ones will work in every context. I mean, you look at the U.N. right now. The U.N. has allowed North Korea to take over for the Council on Disarmament. It's allowed Russia to sit earlier this year on the, on the Security Council. I mean, I'm sorry, but I don't believe there is a [inaudible] body that's going to solve global existential risk concerns.

Moreover, we have past experience with this, and I'll wrap up on this point. I mean, we have tried in the context of chemical and nuclear weapons to negotiate international treaties and controls for many, many, many decades. And the most-- the easiest and best analog here is not in the world of nuclear power-- technology, but rather in bioweapons. In the 1972 BWC, the Biologic Logical Weapons Convention, I mean, everybody signed on to that, including the Soviet Union. And they probably went back home, and they told their scientists to get busy developing chemical weapons, which they did, and develop the biggest chemical weapons stockpile in history. Who else signed onto it? Israel, you know South Africa — they cheated, they didn't pay attention. America probably didn't pay attention either, but we're not, never would to admit it. We have to talk about the realpolitik of AI arms control and regulation. And so much of what's being proposed today is in the realm of hypothetical absurdities. I think we need to be far more concrete and pragmatic about this. We're losing a lot of time talking about silly things like pauses and grandiose AI islands out somewhere in the middle of nowhere. We've got to get more serious and concrete about these proposals and focused on actual real-world harms and not hypothetical things pulled from the pages of sci-fi novels and motion pictures. Thank you.

KORINEK: Thank you, Adam. Now you have described a lot of the difficulties of regulation, and there are many that I sympathize with. You have also advocated self-regulation as a way around. Basically, I think the way you describe the heavy hand of government regulation. How do you see this play out in the space of frontier AI systems right now?

THIERER: Well, we need, we need self-regulation. We're gonna need more than self-regulation. We're going get it. Bottom line is, you're going to have a lot of companies make voluntary commitments. Let's be clear, when companies, when leading tech companies go to the White House or Congress and make voluntary commitments, they're not really all that voluntary, right? This is nice little fun fiction. The reality is there's something more than that. They're going to have some more teeth to it. There's going to be a threat of damage, a sword of Damocles hangs in the rooms. You do this or else, right? And some of that's okay. We've had this in other contexts in the past, and they're going to make some pretty significant concessions, I think, to members of Congress and to the White House. Beyond that, there is going to be sort of coordination at the OSTP level, NTIA level, and many others to deal with a lot of the transparency and explainability stuff. That's gonna be part of the mix. There's going to be an effort by the Federal Trade Commission to enforce AI claims. They've already been very clear in a whole series of different types of reports and blog posts that they're coming after anybody that lies about their AI capabilities. You're already seeing specific agencies get coordinating with the states on things like algorithmic hiring and other types of things.

The bottom line is, it usually doesn't fall to me to be an apologist for the administrative state, but here I am doing it. I mean, the reality is, why not tap all of the regulatory authority and state capacity that already exists in this world before we go layering on new ones? We don't need a whole new layer of, you know, radical, sweeping, horizontal AI regulation when we have context-specific types of things. And then, yes, the best practices are an important part of that. I really do believe they can help form norms and standards among the largest players who are clearly the ones of most importance right now when it comes to frontier models. And a lot of them have already made these agreements and concessions, right? And the question is, how do you go to that next step and get people who are developing internationally, especially things like, you know, Falcon-180B open-source kind of stuff, or whatever China's doing, to agree to the same sorts of things. Because I don't want to go too far beyond that by tying our hands, because that would sacrifice our sovereignty and security as a nation. If we're just saying, "Yeah, they're going to-- we should trust them. We'll do the same thing." No. I'm sorry, I don't.

KORINEK: Okay. Thank you. It's good to see you both as somebody who passionately warns of the dangers of overregulation, but who at the same time acts as an apologist of the administrative state, as you called it, because that's, in fact, the complexity of the issues that we're dealing with, right? So, I want to open it to everybody on the panel now. And I wanted to ask first, Gillian and Dan, to offer their perspectives on the issues that Adam has raised. And then I will hand it back to you, Adam, to basically respond to that conversation. Gillian, please.

HADFIELD: Yeah, thanks. And I want to say I agree with a lot, Adam, of what you're, what you're saying. I think we-- first of all, I agree with you. We have a lot of legal and regulatory infrastructure available that we can be using. You mentioned the FDA, the FTC. We have those. We have tort law. We have, we have some of those tools. I do think that we are-- two problems that I see. One is that we don't have regulatory infrastructure around, and I'm particularly thinking about autonomous systems. So, AI acting sort of autonomously in our, in our markets and in our political social systems. So, I actually see a lack of of regulatory infrastructure there. And so, that's one of the

reasons to say like, you know, a registration requirement is intended to create some infrastructure around that without imposing it until we have better information, the constraint, but it gives us the levers to do that. So, that's one point. The other is that, and I've been writing about this for a long time both before thinking about AI and then in the context of thinking about AI, you know the tools and the levers, the mechanisms we use for that existing regulatory infrastructure, I don't think can keep up with the complexity and the speed and the lack of visibility into AI technologies.

So, I have proposals out there about regulatory markets that I think we need to start building. I think we need to start getting very innovative about how we approach the development of those, of those regulations. So, that's-- so, so, I want to agree with you that we don't-- this is also why I don't think we should immediately be jumping to licensing regimes and going beyond sort of the the development of those standards within labs. But the, the piece that really worries me is the fact that we have such limited government visibility. I think this is the first time in history we have a technology that you can't reproduce in the academic lab. You can't test it in the academic lab. And we have, in private corporations, we have the development of technologies that we've lost visibility into. And I think that's, that's the key challenge facing from a regulatory point of view going forward.

But I, you know, I think we need to be thinking about those risks. I think we need to be imagining how we're [inaudible]. But I agree with you, we do not yet know-- but I don't want to start. You know, I was on our paper about saying let's develop standards, but I think that's a very long process because we don't know. We don't currently know what guardrails we should put in place. And and I certainly would agree with you. We don't want to be going overboard immediately on that.

KORINEK: Thank you, Gillian. We have only a few minutes left. Dan, may I pass you to mic for two minutes? And then let's keep two more for Adam.

HENDRYCKS: Yeah. Great. Yeah, I agree. There's definitely a balance, and we'll need to make sure that coordination and shared standards are incentive-compatible for many actors who we don't completely trust. I think that we all are on the same boat with respect to various catastrophic risks. Bioweapons would end up affecting the entire globe. If we have bots that are able to hack and destroy critical infrastructure that anybody could use in a few years, this would affect everyone. I think also, if we lose control of AI systems, especially militarized AI systems, due to some reliability issues or some access, I think those as well are things that we would want to try to keep the lid on because it would affect the entire globe. So, I think that there are places in which we could work together.

I don't think the situation of-- is as dire as you're painting in terms of the top supercomputers. Those numbers are presumably for supercomputers with CPUs, the US and US companies that really dominate in terms of supercomputers and GPUs. But on CPU's which aren't really that relevant for AI development, I would agree it's a different story. So, I think that I think you're pointing to the fact that other people still are creating a lot of these somewhat powerful, not exactly state-of-the-art powerful AI models, with this compute is a reason to look into your compute governance while we can because this is an actual lever for controlling it. It affects the-- it's an input, but it strongly predicts the outputs of the system, just compute. So, I think if we look at the input of compute to try and regulate that, that could go quite a long way. And we have substantial influence over the GPU supply chain in the U.S. such that it seems like an actual lever.

So, I agree on many of these, on many of these other sorts of risks. it would be difficult to-- a smaller scale risks, not societal scale or civilizational scale risk, would be fairly difficult to coordinate. But I think on some of these we all are on the same boat and hopefully we could, hopefully we could have some shared international precedents or some agreements on that. It'd be worth exploring. We shouldn't assume the worst at the outset because it-- there would be there would be a lot of gains if we can coordinate. So, we have to give that a chance because there does seem to be some room for bargaining there. There is some overlap in interests between us.

KORINEK: Thank you, Dan. Now, Adam, we have, unfortunately, only one minute left, but I would love to hear your perspective on this issue that Gillian has brought up about transparency. Do you think that's maybe, kind of one of the issues on-- one of the low hanging fruits I want to see of something that would be really useful or what is your [inaudible].

THIERER: Yeah, obviously transparency. I'm not going to make an argument against transparency in almost anything. But the reality is, is transparency of what and how. I have a couple of papers dealing with the concept of AI transparency and explainability and the challenges thereof and trying to figure out exactly what we're trying to make more transparent. Because there are obviously some issues there, trade secrets, security concerns, that we have to walk through. But yes, I think a certain amount of that is entirely practical, and I think that's probably where the AI debates are going to end up. But I think contrary to what Gillian did say about limited government visibility or regulatory infrastructure on autonomy, I think the reality is, is that no technology that I can think of throughout history has ever been more thoroughly study or vetted before widespread distribution than artificial intelligence. I really can't name one. By contrast, the Internet really did hit us, you know, out of nowhere. I mean, we really had quite limited government visibility of Internet. I was, I was involved in running in telecom back in 1996, where we barely even mentioned the Internet, right? That's how off-guard it costs.

And so, you know, a ton of thought is going into this. And I'm a big believer in humanity being able to muddle through as they think through problems. And I know it's been tough for the Internet, but the reality is, is that we're putting a lot more thought into AI on autonomy and the dangers thereof, than almost anything I can think of in history. So, I think have a little faith here. And let's not start with radical solutions out of the gate that treat all innovation as guilty until proven innocent. Let's start with the opposite presumption and understand the profound benefits of AI and autonomous systems to move the needle on human progress.

KORINEK: Great. Thank you. So, it seems we have, first of all, some basic consensus about the importance of the topic and about the speed of progress, the depth of dangers that we are facing. And the need for transparency, the need for visibility, so that all decision-makers know some of the basics of what is going on and ultimately getting the oversight and governance, and eventually likely regulation of these systems right, is going to be both an incredibly difficult and an incredibly complex challenge, maybe one of the most challenging regulatory discussions of our time. But I want to thank all our panelists for their contribution to this important debate. And as Adam said at the end, I very much hope that we will somehow muddle through it. Thank you all.