Center for Strategic and International Studies

TRANSCRIPT

Event

"The U.S. Vision for AI Safety: A Conversation with Elizabeth Kelly, Director of the U.S. AI Safety Institute"

DATE

Wednesday, July 31, 2024 at 10:00 a.m. ET

FEATURING

Elizabeth Kelly

Director, United States Artificial Intelligence Safety Institute, NIST, U.S. Department of Commerce

CSIS EXPERTS

Gregory C. Allen

Director, Wadhwani Center for AI and Advanced Technologies, CSIS

Transcript By
Superior Transcriptions LLC
www.superiortranscriptions.com

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES

Gregory C. Allen: Good morning. I'm Gregory Allen, the director of the Wadhwani Center for AI and Advanced Technologies here at the Center for Strategic and International Studies.

AI safety has been one of the most incredible topics – incredibly talked about topics in AI policy for several years now. And we have the perfect person to outline the U.S. vision for AI safety. And that is Elizabeth Kelly, the inaugural director of the U.S. AI Safety Institute, a new institution that was created just in the past few months and is going to be the foundation for setting a lot of the policy, guidance, and other documents around AI safety here in the United States. And also has an important role to play internationally.

Elizabeth Kelly, thank you so much for joining us today.

Elizabeth Kelly: Thank you for having me.

Mr. Allen: We're going to get really into the weeds on what the AI Safety Institute is up to, but before we go there, I wanted to hear a little bit about you and how you came to occupy the role as the inaugural director of the AI Safety Institute.

Ms. Kelly: So, I am thrilled to lead the amazing interdisciplinary team that we have at the U.S. AI Safety Institute. I've spent most of my career in public service, including most recently at the White House National Economic Council, where I was the – one of the driving forces behind the AI executive order, and helped lead a lot of our AI policy engagement globally. Before that, I was in the private sector where I helped start and then scale a startup, from seed stage through acquisition.

I think that this role is really a combination of those two things. It's how do you create a nimble startup that is able to tackle big things, and how do you operate inside a government with all of the complexities, challenges, but also huge opportunities that creates? And that's really what we're doing at the AI Safety Institute, is creating a nimble entity that's able to respond to this incredibly rapidly changing technology, while leveraging the huge power of the U.S. government and governments globally to make sure that we're promoting safe AI innovation.

Mr. Allen: Really exciting. And a really great background, preparing you for this really important moment in AI policy and AI safety. So, I think some folks in the audience will be deep in the weeds in AI safety and have been talking about these issues for years. And others will be just learning that there is such a thing as an AI Safety Institute. So, can you help us understand what is this institution? What problem was it trying

to solve? And what sort of measures and, you know, actions is it going to take to help work through those problems?

Ms. Kelly: So, our mission at the AI Safety Institute is really to advance the science of AI safety. And that means both assessing and mitigating the risks that advanced AI models could pose, so that we're really able to harness the tremendous innovation that we all know is possible, and the benefits that we want to see. You know, we think about our work in a couple of different buckets. You know, one, we are really focused on testing – building out evaluation suites so that we can test new models prior to deployment across a range of risks. And we're working closely with the companies and have commitments from all of the leading frontier model developers to work with them on these tests.

We're also going to be putting out a lot of guidance. We think it's really important to cultivate a robust ecosystem of researchers, academics, inhouse evaluators at the labs, who have clear guidelines on what is best in class, testing and evaluation; look like risk mitigations should be in place and how do we evaluate the efficacy of those risk mitigations. Synthetic content – what are the best-in-class techniques, something that we know is evolving very quickly.

And the last piece is really around research, an unusual function for a government entity to combine testing guidance and research all in one. But I think that that speaks to the fact that there are so many open questions in AI safety right now – how these models work, why they produce some of the outputs they do, what safeguards are most effective. And so we really need to have a virtuous cycle of research, testing, guidance, all informing each other. And we need to be doing it in a way that brings together all of the many stakeholders, both all of the agencies across government for their expertise, countries across the globe, folks in the private sector and academia and civil society, so that we can all work together to get this right.

Mr. Allen: So, you mentioned that you're really focused on advanced AI and the absolute frontier, so not just, you know, the AI that can detect cats in a picture that's existed for more than 10 years now. But really the bleeding edge of the state of the art in terms of AI is where the AI Safe Institute is focused. Is that fair to say?

Ms. Kelly: Yeah, I think that's fair to say. And I think that's in part because of sort of our origin story, as we think about the president's executive order. So as we think about that document, I'm sure lots of your listeners have read it, but I think about it in really three different buckets.

A lot of it is focused on how do we spur AI innovation. And that means

increased money for R&D, attracting and retaining talent, upscaling talent, protecting IT – all of the things that we know we need to have a robust innovation ecosystem.

The second piece is focused on sort of the application and adoption of AI. And as you referenced, lots of traditional machine learning has been in play, in use, for a long time. And we have existing regulators that have authorities to bring to bear, as we think, for example, about using AI in making a lending decision or an employment decision.

What we are really focused on at the AI Safety Institute is the underlying frontier models that are powering all this innovation. And this is something that is so new and so dynamic that we really need to be nimble. And it's why we're focused on the research. It's why we're not a regulatory organization. It's why we're able to work across the government. And I think that's a pretty unique sweet spot.

Mr. Allen: Yeah. The fact that you are not a regulatory agency but you are a safety agency is interesting, I think, because you're not telling companies thou shalt do this; thou shalt do that. But you are saying this is what our consortium has identified, which includes industry participation, academic participation, and your own government inhouse experts has identified as a best practice. Is that sort of fair to say?

Ms. Kelly: I think that's fair to say. And I think that industry has a real incentive to get this right. We know that if something goes wrong, that can be one of the biggest blockers to innovation. And so I think there –

Mr. Allen: And adoption too, I would say.

Ms. Kelly Absolutely. And so I think they're eager to work with us. And we bring to bear really unique expertise across the federal government as you think about the expertise in the defense-intelligence community on cyber that we're marshaling or expertise across Homeland Security and HHS or DOE as we think about different biorisks. So we're really well positioned to help the companies make sure that they are testing and deploying safe models. And hopefully it can be a good partnership.

Mr. Allen: So I want to ask about that partnership, because I think some folks would say, as you said, that industry already has an incentive to be safe, right? They want to be safe because they want their customers to buy more of the stuff. And they probably would if it was safer.

So if industry already has this sort of incentive to pursue safety, what is the argument for why the government needs inhouse capacity around those three areas that you mentioned, as opposed to just leaving it to

industry or academia, et cetera?

Ms. Kelly: So a couple of things I'd say. One is the stakes are so high here. We are seeing AI develop so quickly. I think back to, you know, less than two years ago with the launch of ChatGPT.

Mr. Allen: Feels like a lifetime ago.

Ms. Kelly: It really does. But we're already seeing sort of leaps and bounds in AI innovation with new capabilities, especially in the multimodal space, models that are increasingly scoring higher and higher on different scientific benchmarks used to assess capability. And we know that that pace of innovation is likely to continue. And I think that the U.S. government has a role to play here in helping make sure that its citizens are competent in the safety of the models that are being deployed and put out in the public, that we're preventing the possibility for misuse.

And I think we have unique expertise to bring to bear. As I said, as we're thinking about risks to national security and public safety, this is a pretty core government function. And we at the U.S. AI Safety Institute are really the tip of the spear in terms of marshalling that expertise across government and bringing it to bear with the companies.

The last point I'd make is that because this is moving so quickly you're seeing different standards and best practices coming out of different labs, different research institutions, different nonprofits. There really needs to be a sort of best-in-class gold standard. And we really need to keep raising the bar so that the pace of AI safety, research, and guidance is keeping pace with the development that we're seeing in AI.

Mr. Allen: Yeah. And a lot of folks have talked about this as, you know, can we make it so that the pace of technological progress and managerial understanding of AI safety is increasing at the same rate as AI technology and capabilities. Because AI technology and capabilities are advancing at a frantic pace, and AI safety I think we would hope would keep pace with that.

So a lot of what you just said was outlined in a document that you published, which was the Strategic Vision that came out back in May – sorry, May 21st of this year. You've already talked a little bit about the goals and objectives in that document, but I want to press you on this fundamental research question. You said that it's a little bit unusual for a single government entity to have roles in testing, guidance, and fundamental research. You know, when I think fundamental research, I normally think National Science Foundation. When I think testing, I think agencies like the National Highway Transportation Safety agency,

those sorts of things. So I'm sort of curious, you know, why did you make the strategic choice to actually have government individuals engaged in fundamental research?

Ms. Kelly: So one of the things I would highlight is that we are at the National Institute of Standards and Technology, or NIST. And –

Mr. Allen: A legendary agency for nerds everywhere.

Ms. Kelly: Exactly.

Mr. Allen: But like, you know, this agency has won Nobel Prizes. They're useful in so many different ways, you know.

Ms. Kelly: Right. So, like, this agency's been around since 1905. It's won Nobel Prizes. We have an incredible scientific base, and they have done really good scientific fundamental research over their 120-year history. And so I think this focus on research, testing, guidance all informing each other is part and parcel or why we're at NIST and why we're well-positioned to do this work.

Mr. Allen: Oh, that's interesting. So I think I understand you correctly – and please correct me if I'm wrong – but it's sort of like the fundamental research is sort of aiming at the guidance, right?

Ms. Kelly: Exactly.

Mr. Allen: OK. So you're engaged in this subset of fundamental research that is informing your guidance.

Ms. Kelly: And should inform our testing. I really think, given how quickly this is all moving, testing, research, and guidance should all be informing each other. And I think that you've seen NIST do this successfully in the past, as you think about the work they've been doing on AI in partnership with the community for many years, and how that fed into the AI risk management framework that is used by folks across the globe.

Two other points I'd make is, one, typically government issues guidance and says: OK, great; we're done. We know that we have to be nimble, and iterative, and adoptive; that we can't just issue guidance and let it sit. We're going to have to be updating/revising on a really fast timetable. And we need to be both aware of and helping drive the research so that our guidance is useful and practical for folks in industry. And we've got the team to do it. My head of safety was one of the inventors of reinforcement learning by human feedback, which is one of the key drivers of AI development and AI safety. And we're

recruiting, you know, top researchers from MIT, Stanford, who are working hand-in-hand with the tremendous NIST scientists.

Mr. Allen: And so these are folks who, you know, could be making a gazillion dollars in the private sector, but actually want to contribute to this mission of AI safety, and serve?

Ms. Kelly: Yes. I think it's really inspiring to see the number of patriots in this field.

Mr. Allen: Yeah. And you mentioned Paul Christiano, who was the inventor of reinforcement learning with human feedback. I mean, that's one of the original safety procedures for large language models. So it's not like you've got, you know, government chumps – (laughter) – pretending like they understand this; you've really got some of the leaders in this field already willing to serve in public service, which is great.

So what do you see as the main barriers or challenges to sort of adopting AI safety practices and standards as you're developing them?

Ms. Kelly: So, one, just the fast pace of AI development we know that this field is changing quickly and we have to meet the moment here. Two is the number of open questions that still exist.

We talked about this, you know, why we're seeing certain outputs from AI models. Even some of the developers don't know the answers to these questions and continue to be surprised by the capabilities that are being seen with new models and new advances.

And then, lastly, there's just a coordination challenge, right? We need to make sure that we are working hand in glove with the companies, with civil society, with academia so we're all driving in the right direction and that we're doing it with allies and partners across the globe.

AI safety is a global challenge. It's not like AI development or deployment stops at borders. And so making sure that we have a really coordinated global response, building on the great work that's already happened with the Bletchley Park Summit, the G-7 code of conduct, and Seoul Summit – I could keep going on and on – but I think that's going to be a lot of our challenge and priority.

Mr. Allen: Great. And we're going to get to the international part of the story in a bit but I first want to ask about, you know, you said you want to work hand in glove with industry and academia.

One of the tools that is at your disposal is this February 2024 creation

of an AI safety consortium, a group of 280 companies and organizations that are going to work with the U.S. AI Safety Institute to support its mission.

So what is this consortium? How does it work? How do you benefit from it? How do the members benefit from it?

Ms. Kelly: Sure. So we've got 280 members drawn from civil society, academia, the frontier model labs as well as industry verticals like health care and financial services.

Mr. Allen: How interesting. Yeah.

Ms. Kelly: Yeah, and we've got a pretty great track record at NIST, who has leveraged consortia like this to help inform its work over the past decades. Because this diversity of members we're able to leverage expertise in the different verticals in which we're working.

So we just launched task forces to enable closer coordination on really key questions like biosecurity, like safeguards and risk mitigations, adoption of the AI risk management framework and its recently released generative AI profile, and we're excited to get in the weeds with these folks so that we can improve the evaluations we're developing, improve the guidance we're issuing, and make sure that we're seeing broad community adoption because that's really why we exist.

Mr. Allen: Yeah. That's great. And I think adoption, obviously, matters.

So, you know, some folks who have been following the AI safety debate will be familiar that there was this open letter that had argued that there ought to be a pause in AI research while we figure out AI safety.

Now, you, obviously, were not a signatory to that letter and have not advocated for that but what would you say to the folks who say that, you know, being safe means slowing down AI adoption or slowing down AI research?

Ms. Kelly: We really view safety as driving innovation. My boss, Secretary Raimondo, has said many times that spurring innovation – safe innovation – is our goal at the Department of Commerce and we don't see these two things at odds.

Safety promotes trust, which promotes adoption, which drives innovation, and that's what we are trying to promote at the U.S. AI Safety Institute, and I think, you know, we need to really be doubling

down in terms of the resources that we're putting towards AI safety so we're keeping pace with AI development. But I see them as part and parcel and something that we need to make sure are fueling each other.

Mr. Allen: Yeah. I couldn't agree more.

I mean, I really think about this just as an imperfect analogy. But if you think about, you know, electric cars, if you told me that the electric car was going to have a 10 percent chance of bursting into flames at any given point I think a lot fewer people would buy it or adopt it, right, and if you could make it safer a lot more people would be willing to adopt it, and I think the same is true of artificial intelligence.

The more robust and safe we can make these systems the more markets, companies, individuals are going to be excited to use it for their use case whatever that may be. Great.

So now let's turn to the Biden administration's executive order on artificial intelligence. We just hit the 270-day mark and there was a bunch of deadlines, some of which applied to you, for what federal agencies had to do in response to this executive order.

So can you just help us understand what did the executive order direct the AI Safety Institute to do? What have you done?

Ms. Kelly: On Friday, we released a number of different pieces of guidance. I'll go through the whole list and encourage folks to read the final version online. One, we released final versions of guidance on secure software development framework, a new global engagement plan for standards in AI, and a generative AI profile for the AI risk management framework. We also –

Mr. Allen: So, you guys have been busy. (Laughs.)

Ms. Kelly: We have been busy. Two hundred and seventy days was an aggressive timeline but, as my colleagues at the White House are fond of saying, we are what our record says we are. And there has not been a deadline missed for the entirety of the AI executive order. And we were certainly focused on continuing that tradition and are proud to have done so.

Mr. Allen: (Laughs.) Yeah.

Ms. Kelly: The last thing I'd mentioned is we also released on Friday new draft guidance that provides best practices for developers to prevent the deliberate misuse of advanced AI systems. That's now open for a 45-day comment period. And we're really excited to get the community's

feedback on that.

Mr. Allen: So this is – this is an exciting document. It's a document that I have here.

Ms. Kelly: Indeed! Thanks for that.

Mr. Allen: And have been eagerly reading through it. So the document is titled, Managing Misuse Risk for Dual-Use Foundation Models. And so this is the AI Safety Institute's sort of current best understanding of what – correct me if I'm wrong – the frontier labs really ought to be doing as they are developing and deploying these frontier models. So what misuse risks are you, you know, trying to address in this document? And I'm not going to ask you to go line by line and everything, but sort of what are the steps that you're encouraging – and it is encouraging; this is not a requirement – but what are the steps you're encouraging companies and universities who are developing these systems to do?

Ms. Kelly: Yeah. I won't go through the seven objectives and 21 best practices. I'll let your – (laughter) – exactly. I'll let your viewers read that on their own time. But what I will say is the document is really focused on how do we prevent the deliberate misuse of these systems? And that can mean both the depiction of events that never occurred through synthetic content and the harms that arise from that.

Mr. Allen: Something that's, like, a really hot topic on Capitol Hill right now.

Ms. Kelly: It is, indeed. It can also mean the use by state and nonstate actors to perpetuate more potent and dangerous offensive cyberattacks or enable the development of chemical and biological weapons – risks that we're very closely monitoring and that all of the labs are focused on making sure that we prevent.

Mr. Allen: Yeah, let's talk about cyberattacks, just for one moment. So I am not an especially talented computer programmer. I'm a pretty lousy computer programmer. (Laughter.) But I do remember that, you know, one of my first projects that I did was this game that I wanted to make. And it took me many months, you know, to teach myself everything I needed to know to make this game. And recently, I had the experience of just describing the game to one of the, you know, major commercial, large language models, and it programmed it instantaneously.

Ms. Kelly: Yep.

Mr. Allen: So everything I had spent, you know, months learning how to do, it could just do for me, you know, based on an everyday English kind of a request. So that's lovely for folks who want to make games or want to

make useful applications. It's not great news for a planet that is worried about more and more people being able to perpetrate cyberattacks, right? Because if you think about, you know, describing the capabilities that you want from a system, these large language models – if they're not properly secured, if they don't have the right practices in place – maybe they could lower the barriers to entry for folks who are interested in perpetrating cyberattacks but maybe don't have the technical chops to do it at this current stage.

Ms. Kelly: Exactly.

Mr. Allen: Is that sort of a fair description of the risk you're thinking about?

Ms. Kelly: I think it's a very fair description. And that's why we use the technical term "dual-use foundation models."

Mr. Allen: Because they can be used for good or ill.

Ms. Kelly: Exactly. And that's true both in the cyber context and sort of the coding context, where we're seeing huge efficiencies in terms of coding with the use of Copilot and other tools.

Mr. Allen: Yeah. I recently – I recently spoke to, like, an absolutely world-class program. And this individual said that they've gone from a year ago the systems weren't especially useful to them in their daily work. And just in the past year the amount of progress has been so significant that they now use these systems all day, every day in helping them to generate code. And this is somebody who's at the absolute global state of the art for their domain of computer programming. It's a really incredible pace of progress.

Ms. Kelly: It really is. And I think it's just one indication of the huge possibility that AI offers. The same is true in the bio context, where we're seeing AI lead to new drug discovery and development on a much faster timetable to – you know, it could help solve previously intractable diseases, which is incredibly exciting for all of us. But the flipside of that is that there are real national-security and public-safety risks that could emerge and that we're monitoring as we think about the biological and chemical-weapons space. And so it's the flipside of the coin. And what the guidance really tries to do is sort of outline how can developers be thinking about monitoring and mitigating these risks across the entirety of the lifecycle.

Mr. Allen: Is it fair to say that the document's really aimed at developers?

Ms. Kelly: It is aimed at developers. But I think it's also useful for deployers or adopters. They know what questions to be asking, the companies that are creating these models. We think about that in the U.S. government as, you know, a potential procurer of these systems. And the hope is that it will arm the entire ecosystem to help make sure that developers are accountable, are putting in place safe practices. And that's why the guidance really doubles down on the need for transparency and a lot more publicly available information about what testing is being done, what mitigations are being put in place to really spur this ecosystem of safety.

Mr. Allen: That's really interesting. So, you know, it's not a regulation. You're not saying Microsoft, Google, OpenAI, Anthropic – you know, thou shalt follow this – but, you know, the chief technology officer of a bank might look at this document, and then when they're talking to Google or Microsoft or whomever, they can say are you doing this? You know, show me that you're actually doing this so that I know that my system's going to be safe if I'm going to put it in front of my customers.

Ms. Kelly: Exactly. And again, we all have a role to play here, which is part of why the guidance emphasizes things like mechanisms for reporting incidents so that there is a vehicle for people in the broader community, as they're identifying issues, to let the developers know post-deployment so they can make those adoptions as needed.

Mr. Allen: Great. So that document is out –

Ms. Kelly: It is.

Mr. Allen: – but it's not the final version. You're going to get feedback from at least 280 organizations, you know –

Ms. Kelly: (Laughs.) We hope many more.

Mr. Allen: – who owe you feedback, right. Yeah. And then it'll be the final version of that document. But you said it's going to be updated and all of your guidance is going to be updated on some kind of cadence.

Ms. Kelly: Yes. We'll be finalizing that document later this year. But finalizing is a relative term and that we need to be nimble, iterative and adoptive. And we're really excited to get community feedback on this. We know that what is true now may not be true in six months. There's a lot of great work being done in academia and civil society. And we really want to leverage that best-in-class thinking. And we think we've done a good job, but we know we can do even better.

Mr. Allen: So this is one of your first sort of big deliverables in terms of guidance. Are there any other upcoming deliverables that you could sort of forecast here?

Ms. Kelly: So one of the things that is called for by the president's executive order is additional guidance around synthetic content. So we – NIST released a report at the end of April in draft form talking about sort of the landscape of existing tools and techniques for content allocation, content detection. And later this year we'll be issuing guidance that really points to what are the best practices here and are excited for that, because, again, this is a space that is evolving so quickly. And there's so much attention on it that it's really important.

Mr. Allen: Great. So now I want to shift gears to something else that's in the executive order. So the executive order doesn't do a ton to actually regulate the development of AI. Most of the stuff is, you know, as you're doing here, creation of voluntary guidelines meant to help industry. But there is one requirement that was levied upon industry, which is transparency requirements around not this generation of Frontier AI models but sort of the next generation, what is going to be the state of the art in, you know, four months, six months, 12 months, whenever.

Can you just help us understand what was this transparency requirement that actually called upon the Defense Production Act? And why did the administration choose to go this route?

Ms. Kelly: This goes back to the point I made about the importance of the U.S. government sort of being able to monitor and hence mitigate the risks of AI as we see huge improvements in capabilities with great opportunities, but also a lot of uncertainties. And the requirement that you're referencing is for companies who are training models up to 10 to 26 FLOPs to –

Mr. Allen: Ten to 26 FLOPs, which I'm sure everybody knows. But for, you know – just refresh their memory.

Ms. Kelly: Yes. Basically these are – the way to think about it right now is that most of the models that are in market are trained up to 1025. So if you think about the GPT-4s or others, like, that's sort of the threshold that we've hit.

Mr. Allen: And this is –

Ms. Kelly: This is basically –

Mr. Allen: FLOPs is a measure of, you know, how much computing capacity –

Ms. Kelly: Exactly.

Mr. Allen: – was used to train it. So the current ones are incredibly, unfathomably huge, but the next generation is going to be literally 10 times huger. And they're sort of saying that we – in order to understand who this requirement applies to, we're using as sort of an imperfect but, you know, it-works-for-now metric of how much computing capacity did you use to train your AI model.

Ms. Kelly: Yes. I think you explain it well. And what that says is that these companies have to report to the federal government on the existence of these models, as well as on whatever red team testing that they're doing. Now, this information is kept confidential by the government. It will not be shared more broadly. But the work that we're doing at the AI Safety Institute in terms of what best-in-class testing and red teaming can look like will actually help inform, I think, a lot of the questions that are coming out of this. So, again, we are one government, one Department of Commerce all working together to learn from each other's efforts.

I think we build on that in our guidance by encouraging companies to also make a lot of this information publicly available through transparency recommendations. So, again, we're cultivating this ecosystem of safety because we cannot and are not doing this alone.

Mr. Allen: So you've got this visibility that is, you know, the companies are going to show you sort of what's under the hood of their AI systems; they're going to show you what are their safety procedures, other things like that. And that gives, then, your researchers the opportunity to think about, you know, here's where the current state of the art is; here's where our guidance, therefore, should be.

Now, does that include something like incident reporting or incident monitoring? You know, if these companies have safety – violation's probably not the right word, but if they have safety problems are they then, you know, required to give that information to you, or how does that transparency requirement work?

Ms. Kelly: So this is being led by my colleagues at BIS, and the rules and regulations for this reporting have not yet been finalized.

Mr. Allen: Ah, OK.

Ms. Kelly: So I think there are still some open questions.

Mr. Allen: So what is the – you know, you're at NIST, which is part of the Department of Commerce. Bureau of Industry and Security – BIS – also part of the Department of Commerce.

Ms. Kelly: Exactly.

Mr. Allen: So when I think, you know, who should be in charge of these AI safety/transparency, I sort of naturally think of you. But it is the Bureau of Industry and Security. So what's the relationship for this activity between your two organizations?

Ms. Kelly: We, obviously, work very closely with our colleagues at BIS, but they are the ones who sort of have the regulatory requirement under that DPA authority.

Mr. Allen: Yeah. They own this part of the Defense Production Act, so it kind of has to be them, right? Yeah.

Ms. Kelly: Yes, but we're certainly working very closely with each other and making –

Mr. Allen: You're probably their first call. (Laughs.)

Ms. Kelly: – and making sure that we're learning from each other's efforts.

But I think one of the things that I want to emphasize is that at the AI Safety Institute we are not focused just on models training using 1026 FLOPs, right? We are focused on models that are showing a substantial increase in capabilities.

Mr. Allen: Oh, interesting.

Ms. Kelly: And I think that's really interesting because we're seeing, you know, increasingly less computer-intensive, more efficient models that are actually performing better on a lot of these sort of science-based benchmarks like GPQA or MMLU in ways that I think are continuing to surprise people. So I think it's going to be really important to not get stuck in just sort of how large is the model, but look more strategically about what are the capabilities they possess and what risk they might pose.

Mr. Allen: So now I'm a little bit confused because at the beginning of the conversation you had said that your focus was going to be on the frontier of AI research, these frontier models and the most advanced

systems, but you know, what is the frontier today will be routine, you know, in 10 years. So, you know, what takes a supercomputer today will be on your phone in 10 years. So is the AI Safety Institute's focus on the technological frontier sort of dated from what is the current technological frontier and then anything that's more sophisticated than that forever more? Or are you sort of riding the technological progress wave and always focused on the frontier?

Ms. Kelly: I think we will be riding the technological progress wave. I think the point I'm trying to make is that we are focused on models that are showing new capabilities, and we are seeing, you know, smaller models like the Sonnet model that was released by Anthropic, GPT-4o, many that are actually showing improvements in capabilities. And we want to be monitoring those developments, as well as the developments that are happening as we're seeing more compute-intensive larger models too.

Mr. Allen: Yeah. And I mean, a lot of this guidance will apply –

Ms. Kelly: Exactly.

Mr. Allen: – in mostly the same way, but that is an interesting sort of thing.

Now I want to sort of take us in the direction of international relations and what is the AI Safety Institute's sort of network abroad and relationships abroad. So at the AI Seoul Summit, which I had the privilege of attending and meeting you there, actually –

Ms. Kelly: We did.

Mr. Allen: – in May, 10 countries and the European Union announced the Seoul Declaration, which was a commitment to form an international AI safety institute network for collaboration on AI safety. A bunch of countries that had not previously said they were going to have AI safety institutes like Canada announced that they were going to create these institutions. So this network, you know, why – the U.S. is obviously a party to it, and I'm sure was very influential in negotiating that Seoul declaration. So what were you trying to accomplish here?

Ms. Kelly: As you referenced, we're seeing AI safety institutes pop up across the globe. In Canada, in Kenya, in Japan, in South Korea –

Mr. Allen: Kenya?

Ms. Kelly: Yes.

Mr. Allen: Hadn't heard of that one. That's interesting.

Ms. Kelly: And this is incredibly exciting, and a really important development. But we don't want to be duplicating functions. As I've talked about, you know, the money being spent on AI safety is a fraction of what's being spent on AI development. And it's really important that we learn from each other's efforts and stand on each other's shoulders, so that we're able to work together to really advance the science of AI safety. You think about the great work being done by a number of Canadian researchers on risk mitigation, or the work that a lot of the Singaporeans are leading on synthetic content detection and authentication.

We want this to be a nimble network that is able to learn from each other's efforts, so that we're helping push this frontier of AI safety science. And we want to make sure that we're moving towards, you know, more aligned, interoperable standards and testing so there's a common understanding of what the best practice here is and we're not putting up roadblocks to innovation. Which is what we all want to see.

Mr. Allen: So I realize that this is, you know, relatively young. This agreement only happened in May – you know, a couple months ago. But to the extent that you can sort of say, what is this network going to do, right? You know, you talked about sharing research. Are you going to be organizing conferences? Are you going to, you know, have some kind of bureaucratic secretariat? (Laughter.) You know, to the extent that you can say – and I realize maybe a lot of this hasn't been decided yet – but what is the network going to do?

Ms. Kelly: The network is really intended to, as I joke, bring the nerds together. So we've had, you know, really terrific commitments at the Bletchley Park Summit, at the Seoul Summit from leaders and ministers to –

Mr. Allen: And this was, like, very senior. You know, we're talking Vice President Harris was there, Prime Minister Sunak was there, Prime Minister Meloni of Italy was there. So that was the sort of Bletchley Park, very high profile. You want to go down to the folks who can actually, you know, talk about what it looks like when you're modifying the weights of a, you know, large language model.

Ms. Kelly: Precisely. And I think this is something that NIST has done very well for a long time. And we know the real value that can exist from that. And so our hope is, not only are we exchanging sort of best-in-class research, we're starting to reach alignment in terms of what the best practices for testing and evaluation, risk mitigation look like. Because we're going to see more and more countries not only rolling out safety institutes, but

also putting out legislative regimes. And it's good to have sort of some consensus around what this should look like. I also think that there could be opportunity to do some joint testing, especially as we think about open-source models. So early days, the sky's the limit. But we want this to be decentralized, nimble, and really enable the deep technical conversation between trusted allies and partners.

Mr. Allen: I think that's a really important thing that you just said. So, you know, the range of actors who are participating in this, you know, it ranges from the U.S. AI Safety Institute, which does not have a regulatory function, to the EU AI office, which has this AI safety vertical but is also charged with enforcing the EU AI Act – which is explicitly a regulatory action. It does include regulations that apply to dual-use foundation models, for example. And as I think about it, you have all of these standards. You have all of these, you know, procedures. And sometimes they are going to be recommendations, guidance, and sometimes they are going to be compulsory.

And it would really be helpful to U.S. companies, who really are, you know, the leaders in frontier AI research – it would be really helpful if the standards by which you measure stuff was the same in multiple countries. I think there's a lot of work to be done on interoperability of these different types of measurements.

Ms. Kelly: Absolutely.

Mr. Allen: As I often think about it, you know, you – maybe the European Union will say, you know, we have a lower tolerance for risk. You know, we're going to say that on the safety score you have to be a 10. And the U.S. will say, you know, we have a higher tolerance for risk. On the safety score you only have to be a five. But do we agree what a 10 and a five are? They are they measured the same way in both places? Because if they're not measured the same way in both places, then every U.S. company is going to have to develop, you know, different compliance departments and different scientific and technical underpinnings to support those compliance departments for all these different countries around the world. I mean, would you agree? Is that a fair description of what the point of this collaboration can bring?

Ms. Kelly: I think it's one of the outcomes that we want to see, is really moving towards global alignment on what safety should look like, so that we're able to unleash the tremendous innovation that we want to see. The U.S. is the clear leader in AI innovation. And we are proud to be the host of the companies and researchers that are really on the cutting edge of this work. And we think we can be an equal leader in AI safety and that each of them will support each other.

Mr. Allen: Yeah. I think that interoperability, you know, point for me is really strong in terms of even if the United States does not go down, you know, a strong regulatory path, the way the European Union has, you really want interoperability of those – of those standards. And I think this is also a vehicle for U.S. leadership on this area, right? Do we want – like, this work has going to be done somewhere. Do you want the U.S. in the lead, you know, setting the pace of the conversation? Or do you want us to be reacting to what's happening around the world?

Ms. Kelly: I think that, as Secretary Raimondo has said, we need to be leading on AI safety as well as AI innovation. It's part of why we're working so closely with our colleagues at the State Department and across the government. And I really think that this network is going to be key to that, but also help boost capacity in a lot of the member countries and make sure that we're learning from each other.

Mr. Allen: So, you know, one thing that's always helpful, at least in my own understanding of these issues, is, like, an analogy. So this international AI safety institute network, is there other – is there some other technology or some other precedent that is in your mind as you think to, like, what you're trying to accomplish?

A lot of folks talk about, for example, the International Atomic Energy Agency. I don't think that's a great fit for the AI situation. There's also the International Civil Aviation Organization, which deals with, like, commercial airlines and that sort of regulation and standards. Is there a precedent in your mind that is helpful as you think through, like, what the international AI safety institute network should be trying to accomplish?

Ms. Kelly: I think we are learning from a lot of the different models and organizations that exist. And have been, you know, fortunate to sit down with people who are running those organization and learned what's worked, what hasn't worked. I think that our real takeaways have been that we need to be decentralized. We need to be nimble. And those would be a lot of the key principles that we're bringing to bear as we think about standing up this network.

And I think there are many contexts where the U.S. has been able to lead in safety in order to fuel innovation. We think about, for example, airplane safety requirements, which have been adopted across the globe and really reflect a lot of U.S. leadership and know-how here. Now, obviously, you know, ours are not requirements. They're guidance and best practices. But a lot of the same principles can apply.

Mr. Allen: That's great. So you've got this international network. But folks who've been, you know, following my own work, especially a lot at the G-7, know that there's other stuff going on in the international sphere when it comes to AI. So you've got the work going on at the G-7. There's also some work going on at the G-20. There's work going on at the OECD. There's work going on at the United Nations – all sort of targeting AI governance, many of those conversations talking about AI safety to some greater or lesser extent.

And so I'm curious you know, how you view the AI safety network as distinguishing itself, and how it fits into this, you know, landscape, which you could argue is sort of crowded at this point, right, with international institutions trying to have a say, and have a voice, and make an impact on AI.

Ms. Kelly: Our goal is really to feed into and inform all of those conversations that are happening. You know, conversations where the U.S. has been a leader. If we look at the G-7 code of conduct, that has a lot of similarity with the White House voluntary commitments that were adopted last July. We're proud of the U.N. General Assembly's adoption of the U.S.-led Declaration on AI. And I think that the work that we're doing in terms of really bringing the technical experts, the researchers, the folks who are hands on keyboards together to figure out what best in class is here, to move towards aligned, interoperable standards, is going to be really powerful and helpful for all those other efforts that are happening.

Mr. Allen: That was very well put. And I think the sort of technical expert dimension is something that I do think is distinguishing. I would also say, you know, having been personally involved with a lot of the G-7 work, I do think that you are, in some ways, an implementation of a lot of the recommendations of some of the G-7 work of, you know, we've got the bureaucrats together, now let's get the actual technical wonks together to start hammering out some of these standards. You know, interoperability – which I've been harping on too much in this conversation – that was one of the recommendations of the G-7 in 2023 and again in 2024. And I think this network could be one of the mechanisms whereby that interoperability really comes to pass.

So you mentioned – you know, is it fair to say in your three big areas of, you know, fundamental research, testing, and guidance – are all three of those areas going to have big international components? Or is one of them more international than the others?

Ms. Kelly: I think it's still early days. And we are a quickly evolving, growing startup.

Mr. Allen: You've already had 60 days. Why haven't you – I'm kidding, no. (Laughs.)

Ms. Kelly: Exactly. No, I think that we want to work with allies and partners across all of these. And I think the network is a great way to achieve that, both in terms of sharing research and building on the research collaborations that NIST has done for many years as well as the, you know, partnerships we've announced with the U.K., AI Safety Institute through the MOU with the EU, with the dialogue that we're having under the Trade and Technology Council to help inform each of our efforts.

And we want to move towards, you know, more aligned interoperable guidance and standards. So I think it'll be informing all of it and hopefully providing an opportunity for U.S. leadership here.

Mr. Allen: That's great.

Now, I want to ask you about that MOU you just mentioned with the U.K. AI Safety Institute. So the U.K., obviously, with the Bletchley Park Summit was the first to announce that they were going to create an AI safety institute.

They have a lot of resources. My understanding is that they have a budget of 50 million pounds per year, which is around $65 million U.S. They've got dozens of technical staff, you know, already hired and working for them.

And so, you know, what has happened with this MOU? What are you doing in partnership with the U.K.? How's it been going so far?

Ms. Kelly: So far so good, and I'm optimistic that will continue. We, you know, work closely with our U.K. partners and have been able to leverage a lot of the great work that they have done, for example, building out evaluation suites.

They actually open source a lot of this work through their Inspect platform, and this really goes back to the idea of how do we all learn from each other and leverage the work that is happening so that we're not reinventing the wheel, not duplicating, and that's part of what this partnership allows us to do.

Mr. Allen: Great.

So the U.K. AI Safety Summit was borne out of the Bletchley Park Summit. Then we've already talked a little bit about the Seoul Declaration, and there's going to be another successor to the U.K. AI Safety Summit, which is going to be in France in February.

And so I'm curious, you know, the international collaboration of AI safety, do you see it as continuing to take place at these summits? But you've already got this convening happening in November. So how is all this going to fit together?

Ms. Kelly: So our convening in November is, as I said, really focused on bringing the technical experts together. If we look at Bletchley Park, at Seoul, and as what's planned for the AI Action Summit in Paris in February these are really about getting the commitments from leaders and ministers to prioritize AI innovation and AI safety.

But our hope is that the work we're doing can really be about sort of the implementation and adoption of that, going a level deeper as well as informing a lot of what will come out of that. The AI Action Summit in Paris has a whole track around AI safety. And we want to make sure that we are, you know, informing and feeding into a lot of that work, and think that the timing of that, the folks that we're bringing together at a more technical level will be a really good collaboration.

Mr. Allen: That's great.

So you've got a ton on your plate. You've been assigned a lot of homework by the Biden administration's AI executive order. You've already got some deliverables out the door. Can you just sort of help us understand what should we expect from the AI Safety Institute over the next, say, 12 months?

Ms. Kelly: One thing I really emphasize is we are excited to begin testing of frontier models prior to deployment, and I think we're in a good position to begin that testing in the months ahead because of the commitments that we've gotten from the leading companies towards us.

Mr. Allen: The leading companies have said we want the U.S. government to do their own independent testing of our systems. They've requested this from you.

Ms. Kelly: It's part of the commitments they've made with the White House voluntary commitments, the G-7, and I think that we're all excited to

work together on this. So I think –

Mr. Allen: You know, when the – sorry to interrupt, but you know, when the – when the White House voluntary commitments, you know, came out, I was sort of like, but who in the U.S. government is actually qualified to test, you know, what these companies are doing?

Ms. Kelly: (Laughs.)

Mr. Allen: And now we actually have technical experts who are qualified to –

Ms. Kelly: Exactly. We have the person who invented reinforcement learning by human feedback. You know, we've got teams of experts coming out of a lot of the leading safety labs. And so we're excited to roll that out. I think we're also excited –

Mr. Allen: And that's going to happen this year.

Ms. Kelly: Mmm hmm.

Mr. Allen: Great.

Ms. Kelly: I think we're also excited for the guidance that we'll be releasing. Obviously, we're going to be finalizing that guidance on potential misuse of dual-use foundation models and updating it to reflect all to the community feedback and input we're excited to get, as well as putting out guidance on synthetic content tools and techniques later this year.

Mr. Allen: That's great. Anything else you want to highlight?

Ms. Kelly: We're really excited for the launch of the AI safety institute network –

Mr. Allen: Yeah.

Ms. Kelly: – and for this convening in November to bring together, you know, not just the allies and partners who have stood up safety institutes that are similar entities for the technical conversations on things like benchmarks and capabilities, risk mitigations, but also to bring together the broader civil society, academia, industry, technical experts. And I think the fact that we're hosting in San Francisco really speaks to the U.S. leadership role here, and how we want to continue to maintain and grow that.

Mr. Allen: Well, it's been really encouraging not just that your organization now exists, but that it's had such a fruitful, you know, first few months of

existence and is already, you know, producing work that's really going to move the needle.

Elizabeth Kelly, thank you so much for spending the morning at CSIS.

Ms. Kelly: My pleasure. Thanks for having me, Greg.

Mr. Allen: Well, this concludes our event. And if you'd like to watch it again, you can visit our website at CSIS.org. Thanks.

(END.)