

THE BROOKINGS INSTITUTION

WEBINAR

HOW CHINA AND THE US WILL REGULATE AI MODELS FOR SAFETY

Tuesday, October 6, 2026

This is an automated transcript that has been minimally reviewed. Please check against the recording for accuracy. If you find any significant errors of substance, please let us know at events@brookings.edu.

PANEL

MODERATOR: MARK MACCARTHY

Nonresident Senior Fellow, Center for Technology Innovation, Brookings

KYLE CHAN

Fellow, John L. Thornton China Center, Brookings

JAMES ANDREW LEWIS

Distinguished Fellow, Center for European Policy Analysis (CEPA)

MATT SHEEHAN

Senior Fellow, Asia Program, and Co-Director, China AI Initiative,
Carnegie Endowment for International Peace

* * * * *

MACCARTHY: Welcome, all. I'm Mark MacCarthy. I'm with Georgetown University and the Center for Technology Innovation here at the Brookings Institution. And I'm delighted you all can join us for this panel discussion on how China and the US will regulate AI models for safety. We've got a wonderful panel to discuss this stuff, and we'll be going for about an hour, and we have some time for your questions in the last fifteen minutes or so.

So, with that intro, let's get started. As many of you probably know, over the summer AI company experiments resulted in a number of disturbing events, including cybersecurity attacks on third-party websites. And there are other unexpected and potentially harmful actions. The most famous was the cybersecurity attack on Hugging Face.

But Axios recently reported that there may be as many as tens of thousands of these incidents that the companies are investigating. And some of these incidents involved problematic access to government websites in the United States and in Australia. And so of, of course, this fueled lots of media speculation about out-of-control AI agents causing catastrophic damage.

Anthropic says they want a coordinated pause in AI development to avoid these risks. President Trump, however, says that these existential risks are a hoax, and he rejected any attempt to slow down AI development. He wants to step on the gas to win the race with China. Whoever wins AI, he says, wins. But he's also appointed an AI czar, Jay Clayton, the current director of national intelligence, and he's formed an AI task force that's going to be producing a report in 120 days among other things on AI safety recommendations.

And the companies themselves have taken some steps to promote safety. OpenAI is delaying the release of one of its latest models, and it's paused training of another one after it escaped its computer sandbox. And some law, lawmakers are beginning to offer their own attempts to address AI safety issues including the idea of having companies install a kill switch on the models and requiring independent safety assessments of the models. And in one case, there's an attempt to ban any attempt to achieve super intelligence.

So this was the context that really surrounded the meeting between President Trump and President Xi in September where they discussed, among other things, how to work together on AI safety. And what came out of that meeting was an agreement to have a future meeting, which is a probably a good first step, and that may take place as soon as November. And the two countries apparently want to set up a communications channel to report AI incidents that might have national security implications.

So we're gonna get down into all these issues with an expert panel. Today we have Kyle Chan, who's a fellow in foreign policy here in the John L. Thornton China Center at Brookings. We have Matt Sheehan, who's a senior fellow with the Asia program and co-director of the China AI Initiative at the Carnegie Endowment for International Peace. And James Lewis, he's a distinguished fellow at the Center for European Policy Analysis. The bad news is that we're gonna be missing Kendra Schaefer who's a director of tech policy research at the business strategy firm Trivium China. She's under the weather today. It's that part of the year. So we will miss her insights and expertise, but we have a good panel to get into this topic of how the US and China can cooperate to regulate AI models for safety.

But we'll start first, and I'm gonna throw this question to Matt initially to talk a little bit about the AI marketplace in China. This will help give us some context for the possibilities of

cooperation. We know that the Chinese government and the companies really are focused on AI diffusion. Let's try to get this stuff moving through the economy. They're less focused on the idea of developing super intelligence. And I guess the key thing is that open weight models are apparently a key element of that strategy. So why don't you tell us a little bit about how that's going and how that aim to distribute these new AI models throughout the society and the economy. How's it going in China?

SHEEHAN: Sure thing. Thanks very much for having us, Mark, and the whole team. So, yeah, I think you laid out part of that correctly in the, the focus on diffusion among the Chinese government. So essentially plugging AI into the economy. The name of the major policy initiative that they've announced on this front is called the AI Plus plan, where AI Plus means AI plus manufacturing, AI plus science, AI plus healthcare. And this is a pretty standard playbook for how China tries to diffuse information technology into what they would call the real economy, which is basically sort of more tangible physical goods and to a certain extent, services. So that has been the government's sort of big bet on AI, that it is a general purpose technology and that they can use it to upgrade their economy, you know, leap over the middle income trap, climb global value chains and, you know, improve livelihoods for people in China.

I think it's worth distinguishing that governmental bet, party, you know, CCP bet from the, what's driving the frontier labs, because I do think that while they don't have sort of the long history and the, like, the laser focus on, you know, AGI super intelligence and then sort of the risks associated with that, they don't have that long history as the, some of the U.S. labs do, where that's been like in their founding DNA since the very beginning. But I do think those labs are generally pushing for some version of AGI for now they're talking more about what's called recursive self-improvement. So models that build the successor model, which many people think is sort of the takeoff point for an intelligence explosion of various kinds.

And so I wouldn't say that the government and the frontier labs are sort of at odds over this. In some ways it's like a division of labor. You know, really, you know, Zhipu, z.ai or Moonshot, these startup labs, they're not really the ones who are going to be building the sort of AI into manufacturing interfaces. Their job is to build the models and then it's the job of sort of other companies, in some cases state-owned enterprises, in some cases robotics companies, to do this kind of connection between the frontier model and the economic diffusion that they're aiming for. So I think that's one element of it.

How it's going, it's very hard to tell on the diffusion stuff. You know, there are lots of reports coming out of China of, you know, all the great upgrading that's happening and how it's boosting this and boosting that. You know, we don't see it in the macroeconomic sort of productivity numbers yet, but maybe that's not to be expected at this point in time.

I think in the past, China has been very good at this type of diffusion. I sort of-- I lived in China in the 2010s, and I sort of watched this play out with the mobile internet, where, you know, it starts from just, you know, fun and games on your phone, and next thing you know, it's street vendors doing digital payments. It's all kinds of sort of online to offline delivery services. You know, a real kind of plugging in of the mobile internet into the physical world. And China has been good at that across also its manufacturing sector. I think part of the issue with AI is it's significantly more sort of technically complex to plug AI into industries like manufacturing, and it costs money.

And whereas in the past, the local governments throughout China were sort of flush with money that they could use to subsidize this kind of stuff, right now, local governments in China are not flush with money. They're very cash-strapped. So that, to me, is one of the big questions that's gonna play out in the, you know, sort of next year or two is, as the frontier models, the companies continue to stay probably somewhere close to the frontier, and as the government is pushing for diffusion, is that sort of transfer into the broader economy happening, or is it reliant on sort of a, an old China model, an old model of the Chinese sort of political economy that doesn't quite apply as much anymore? I don't feel like I have a good sense of that just yet, but it's something I'm watching.

MACCARTHY: So the, you know, the, I think you're focusing on a problem that exists here as well, which is how to get from the models themselves, even the models with appropriate sorts of harnesses and orchestrations into the real world of productivity improvements in enterprises. We're struggling with that as well, and it sounds like China has the same fundamental business and technology problem.

But they seem to be closing the gap, and Kyle, you may wanna jump in here. You know, Bloomberg reported the other day that some of the Chinese models are only 3% off the performance peak of the best US models and that's down from 9% off the peak in May, and 15% earlier in the year. What's the significance of this closing of the gap between the best Chinese models and the best US models?

CHAN: Yeah, great question. So the gap depends on the day of the week, the latest model releases coming out of China, the latest model releases coming out of the US, and it's hard to really pin down what that gap is and how it's fluctuating over time.

So yeah, one can look at some indicators and compare, say key benchmarks for US publicly released models versus Chinese models like from DeepSeek or Zhipu or Moonshot, which is the maker of the Kimi K3 model and see that on the benchmarks they look pretty close. And then you can look at other sort of indicators like, for example the UK, yeah, well, the UK AI Safety Institute did a test where they found that Kimi K3 was actually quite quite impressive. And Anthropic recently came out with a test also saying GLM 5.3 from Zhipu was also quite impressive and maybe only a few months behind some of Anthropic's best models. At the same time though, there's models that have not yet been released by the US labs that may be even better, so it's hard to gauge.

But what I think this means is that for a lot of tasks, especially for regular individual and business users the Chinese models are very good and likely quite adequate for a wide range of capabilities, and especially that includes these more recent agentic tasks that have really been taking up a lot of people's attention and taking up a lot of the, the AI usage in general.

There is still that question about what that remaining gap means, and that can mean, for example, different capabilities for cyber, whether that's for cyber attack or cyber defense. It could also mean different capabilities for the cutting edge of applying AI for science and research and development. So it could still mean something. But then, yeah, I guess the question, and it comes back to what Matt was saying is, you know, how much does that capabilities gap still matter? And how much does the diffusion part matter and the increasing integration into people's daily lives? And even this I would say that this changes constantly.

Earlier this year, I would have argued that China was moving faster on agentic AI in sort of personal and business use with the boom in OpenClaw. So OpenClaw is this sort of open source agentic harness that allows you to connect your personal computer and your personal files to, you know, an AI system and have it do real things for you. And there was a sort of OpenClaw craze earlier this year. And now actually what's interesting is at the time there was some enthusiasm in the US but maybe also some caution. And now we see sort of more polished versions like Meta's Muse that are trying to do similar things but in a sort of more neatly packaged application for users. And that's taking off, I think, faster in the US now arguably than in China.

So it sort of depends. Even on the diffusion part, that depends on the day of the week. But I think ultimately these, these close capabilities mean that a lot of the issues a lot of the benefits and the issues that we face in the US China's facing as well, and the rest of the world now has to kinda contend with really powerful models from both countries.

MACCARTHY: Yeah. There, I mean, it's been a while where people have been saying that the Chinese models are behind this but yeah, they're, they're good enough for everyday use. But now they seem to be moving closer towards the frontier as well and perhaps challenging US models there as well. And there's another area where the, the Chinese companies seem to be making progress, and Jim, I'm gonna ask you to jump in here.

China's companies appear to be making progress in, in making and producing their own AI chips for training and using their AI models, and this is despite the restrictions on access to US chips and high-end manufacturing technology. What are the developments there and what's the significance of China closing that gap as well?

LEWIS: Thanks, Mark. And let me start by saying I've been in about eight major negotiations with the Chinese on arms control. This just doesn't feel like the time to start. So we'll-- I know we'll get to that later, but the Chinese goals are largely the same, to protect the party and to catch up with the US. And so chips have become a central focus for them.

Xi Jinping, by the way, has said the same thing as Donald Trump, "AI is the future. Whoever wins, it'll be the leader." And they have put a lot of effort. In some ways, our own export controls have helped them because by cutting them off, we've incentivized them to build their own. So the-- it's not like there's a single chip that will rule them all, right? It's the combination of things. And the Chinese are good. They're not as good as us, but they're good.

They're also, I think, not ready to negotiate. I was surprised that they didn't ask for more in the summit on relaxing the technology restrictions. They certainly didn't get anything. That's clearly a card the US has. But we're not gonna, just as within the-- and I know I hate, everyone hates Cold War analogies, but just as the Soviets lagged behind in certain areas, the Chinese lag behind. That doesn't affect the chances or the outcomes of negotiation. And right now, I'd say the chances of a successful negotiation, even on something as dumb as a hotline are really small.

That could change if the Chinese improve capabilities. I don't-- Maybe my colleagues would know better, but what I hear from people in China is they don't have quite the same existential fears that we have. You were saying before, Mark, that might be changing. Could be. But at least now they're not-- they don't look under their bed every night for Terminator.

MACCARTHY: Well, maybe they should look in their mobile phones instead and see if it's lurking there. So, you know, Kyle, maybe you should jump into this. You know, for years the Chinese authorities focused their regulatory efforts, AI regulatory efforts on content control. And so always what they do, they get a new technology, let's make sure that it's not used to say or spread bad things. They're moving a little bit towards protecting kids, okay. But they haven't really focused on the kind of safety issues that have obsessed US policymakers and researchers, the cybersecurity threats, the bio threats, the loss of human control.

As Jim suggested, you know, I've picked up some indications that's changing but maybe, Kyle, you could bring us up to date on what you think about that. And then, Matt, I'm gonna ask you to jump in as well on, on that.

CHAN: Yeah. So AI presents as I mentioned earlier, a wide range of benefits, but also a wide range of risks. And it's, you know, the comparison is often made to, like, nuclear weapons, and that's, like, a very specific risk, and I think we're all quite on the same page. And back then, the US and Soviet Union were pretty much on the same page in terms of what that meant in terms of a risk. But AI actually involves a whole range of risks: social, political for China as well as cyber, the technology itself, and then existential risk.

So just to start at the top, on the existential risk side, that is one of the biggest differences I see right now, at least between Chinese policymakers, the Chinese AI community more broadly, versus the US AI community, where especially most recently, we had that one Anthropic employee who quit and said that AI will kill us all within the end of the decade.

That, that kind of fear of a P doom, a probability that AI could wipe out humanity, is something that has been talked about debated, discussed and really motivates a lot of people in Silicon Valley and has been part of sort of the American AI culture for quite a long time. And so when people say within the American AI labs say that they worry about these things, you know, I take them seriously in terms of their, the authenticity of their beliefs in this and their, the genuineness of their concerns. And then I think now more recently we see the American public worrying about some of these same things.

But I see something quite different happening in China and in particular among Chinese policymakers, which at the end of the day, they're the ones who are gonna have to regulate and who are currently regulating the space. And for them you really hardly hear about discussions of existential risk. There is one standards body that has noted this but they're I would argue somewhat of an outlier. For the most part, when you look at statements, speeches by political leaders it's usually about different kinds of risks.

And to your point, Mark, what are those risks? A lot of those risks that China was focused on earlier were around content, especially applying a kind of social media or internet control playbook to AI and generative AI in particular, right? So they didn't want, they wanted to regulate deepfakes. They want to regulate the answers that these chatbots give to certain sensitive political questions in China.

But then more recently, I would argue, we're seeing a shift not just in terms of what risks China is focused on, but where in the party state apparatus we see those risks being expressed. And I would argue that we're seeing a greater focus and interest in things like cybersecurity risks especially in a post-Mythos time when people in China are watching very

closely what's happening in the US and there was a lot of commentary around the cyber capabilities, in particular of Anthropic's Mythos model when it was first announced.

And then the other major area that I would say there's increased focus in China is on autonomous systems. And even though the terms are different in Chinese and there's a lot of debate about what they mean, but generally speaking, the concept of loss of control and not having full control over AI agents as they run around and do things you know, maybe in swarms of hundreds or thousands there's increasing efforts to regulate or at least warn about some of those risks.

So that's sort of where I see the, the broader hierarchy of risks in China and where they're sort of moving up the ladder in terms of what they're focusing on.

MACCARTHY: Yeah I think that's a great summary of some of the changes that are beginning to move through there. But Matt, I wanna ask you to focus on a slightly different issue, which is the risks of open weight models.

China, most of their models are open weight, which means once they're out there in the world, they can be changed and adjusted, apparently stripping away many of the safety features that are built in with minimal time and compute. And yet China, you know, is aware now that these open weight models, as they become increasingly capable, could fall into the hands of non-state actors and could create some risks to China and risks to the United States obviously as well.

What's their thinking about these open weight models? Is there some sort of idea that maybe there should be some gating of the open weight models, that they should they should go through some kind of review before they're more generally available? What's the current state of play in that area?

SHEEHAN: Sure thing. Yeah China's views on open weight sometimes they'll call it open source and also refer to open weight, have evolved over time. And I think they essentially sort of stumbled in, at least from the government's perspective, sort of stumbled into what they have felt was a good situation for a while of their labs primarily releasing open weight models and that being a benefit on a few fronts.

I say stumbled into it because it wasn't-- it's not that at the very beginning, you know, the government was telling all Chinese labs to go open weight. Some were closed, some were open. The government had its concerns about open weight models, but primarily because they saw open weight models as a vector for US models to essentially infiltrate the Chinese market and spread, you know, what they would call misinformation.

So initially they were a bit defensive about open weight models because they were worried about, for example Meta's Llama model. Then more Chinese models started going open weight, I think in large part due to the, essentially the, you know, the personal views of some of the founders and also some business incentives for the companies.

A lot of the founders in China are pretty-- lean pretty hard into open weight as a, kind of as an ideology on a certain level. And it makes sense from a business perspective. You know, if you're a Chinese company that is, you know, unknown outside of China and even really inside of China, and you know that in the West there's probably going to be some level of

suspicion or resistance to the idea of just, you know, downloading a Chinese AI app or using a Chinese API, then releasing your model open weight and allowing the whole world, the technical world, to play with it, to inspect it to implement it, is a way to sort of gain attention, gain at least usage share, if not like revenue, because they don't always gain revenue from the open weight models.

So I think it's you know, the government starts off unsure about this. Companies kind of, for a variety of reasons, stumble into these open weight releases. And then because DeepSeek did it, kind of all the other companies felt pressured into doing it. And we got to this status quo that I think from the Chinese government's perspective, at least to date, has been pretty good. You know, it feeds into their international messaging on AI. Their messaging has always been, you know, the United States is the technological hegemon that wants to keep this tech just for itself, and we are the open, development-focused, Global South-friendly country that wants to share. That's been the message. And I think the open weight models essentially gave a certain amount of reality to what was previously just rhetoric, at least it made that possible.

It also helps with, as we were talking earlier about diffusion. You know, open weight models are just better or can accelerate the diffusion of the technology into the economy because you do not need to constantly be paying essentially a fee, you know, a subscription fee to the frontier AI companies. You can implement it and tweak it locally for your own purposes. So that's kind of maybe gets us up to like 2025, 2026.

Then, you know, for a while in the West, in the United States, there's been this fear of open weight models. You can strip away the safeguards, and they can be used for cyberattacks, bio, et cetera. China has been essentially unconvinced of that for a while, but starting around, I'm trying to think of when exactly to pinpoint. I'd say it significantly picked up in around July with the release of Kimi K3 and then sort of subsequent documents released by the key regulators, where they start to just slightly hedge on this sort of you know, open weight absolutism.

They say, "Well, maybe there are certain capabilities that need to be you know, checked first and then released. Maybe there are certain capabilities that should not be released." We haven't seen this translated into a sort of a hard regulatory action in a way, but the, I'd say the conversation is moving in that direction.

And I guess just the last thing that I would note on this is I think in many ways the recent spate of, you know, model rogue agents, model escapes, and a variety of other news coming out regarding the US models has somewhat muddied this picture, and I think it's tweaked my own views on this in that we're finding that no matter how well, you know, the US closed models try to implement safeguards, the models can still be jailbroken. It might be easier to jailbreak a US closed model than it is to download and remove the safeguards of a Chinese model from a technical sort of capabilities perspective. And so at the same time that China seems to be inching closer to some level of concern about what happens with open weight models, there is a bit of this sort of muddying of this, what I think was for a long time maybe an overly black and white binary of, you know, closed safe, open dangerous.

We'll see. Events, you know, just this morning, South Korean banks were attacked by Chinese, allegedly by Chinese models. At least that's how some are reporting it. And incidents like that in South Korea, in China, they could change this calculus quickly.

MACCARTHY: So it's a moving landscape and I think that's an important element as we come to later on talk about the discussions between the United States and China. It's not as though everyone's got a fixed position and that, you know, everything is moving on both sides.

And then Jim, I wanted to switch to you on this question of open weight models. There was a lot of talk over the summer in the United States about shutting down Chinese open weight models in the United States, despite their popularity with U.S. enterprises. There partly was a fear of distillation, partly it was just that we didn't want the competition. What, what's going on there?

And there are other areas where the U.S. is trying to, you know, do things that would constrain China in various ways. The, that remote server loophole that is being discussed in Congress is also out there. And the MATCH Act which is in, in the National Defense Authorization Act looks as though if it's got a chance of actually becoming law. And it would ratchet up the controls on lithography equipment which can be used to make AI models.

What's what's going on with all those changes in the U.S. side, on constraints on the ability of Chinese to develop and use their models?

LEWIS: Well, well, thanks, Mark. And first of all, this is a political question, it's not a technical question. And we don't have the pieces for serious negotiation. Think of the stuff you need to start a serious negotiation. We don't have it on either side, so that kind of limits it.

Having read the Match Act, the first thing that occurred to me, of course, was that if you want to ask for favors from your allies, maybe you should be nice to them beforehand. So that's the drawback, right? We're coming and saying, "You know, we're bailing out of NATO. We love Russia, but by the way, can you help us on China?" That would be a hard one to sell. So I think that you've got suspicion on the Chinese side and in the US too, that a lot of the fuss over this is to preserve America's technological lead, and more importantly, to preserve the lead of two companies in particular.

And there's a lot of suspicion that what's driving this is self-interest as much as any regard for the future of humanity. And in the cases of the few Chinese I've talked to about this, I said, "So what do you think about, you know, Anthropic or OpenAI?" And they all, their letters always, their answers always have the same three letters in it, IPO. So this is seen as much as a way, and I tend to agree with them. This is as much a way to stifle competition as it is to hold back the dread menace of AI. And that's not a good way to start negotiations.

But there's so many pieces we're missing. You need presidential support. We don't have that. You need data behind it. Maybe we'll get that with NVIDIA's developments that will track these problems. Maybe, but that will be months away. When we negotiated the cyber agreements, and Ryan Hass at Brookings was of course fundamental in that. It took four years to get the Chinese to agree to something in a small group, actually bilaterally, and it took seventeen years to get the UN to agree. So don't hold your breath, and it doesn't help when the issues of self-interest cloud so many of the proposals we see now.

MACCARTHY: So Jim, before we get into the, you know, prospects of cooperation between the U.S. and China, I wanted to get your take and any, you know, Matt or Kyle, if you wanna jump in, that would be good too.

What's your take on the latest proposals from the administration on AI safety, this AI czar and the report, and the safety principles? Is this a, a move towards real regulation, or is it just sort of hand-waving and you do-it-yourself companies? What, what's going on there? What's your assessment?

LEWIS: It's a mixture of both, and I think one thing that I would agree with President Trump on is regulation's probably a bad idea. The idea that you'll have a third-party review, we did that in cybersecurity, and here's sort of the bottom line. Six months, a million dollars.

So any company that can't afford to take six months off offline and spend a million dollars for outside review, you're not gonna get reviewed. It doesn't work. There are alternative models for regulation, and this sort of approach could lead to that, but it's a start. It's as important to avoid the pitfalls of regulation as it is to actually think.

And it gets back to the data question. When you negotiate an arms control, you can usually say specifically, "This is what I'm worried about. This is what I want. I want a five hundred-mile range limit on missiles. I want a kiloton range. I want agreement that there are certain targets that will not be attacked outside of the Geneva Convention." If you can't make those specific proposals, and I haven't seen any, then it's kind of a waste.

And so as a starting point, what the administration is doing is good. If it continues to move forward and there are alternative regulatory models if, I was thinking last night the way CFIUS regulates foreign investment might be a better model for AI than some of the other things we do. If they continue to move forward and allegedly what, they have three months to come up with a report, 90 days, something like that, this new safety board. Let's see what they write then. So good first step if you keep moving.

MACCARTHY: Got it. Any comments from you, Matt or Kyle on, on where we are in the United States?

SHEEHAN: I think one comment on sort of regulation, over-regulation especially, you know, a lot of this is seen as sort of, with suspicion as like regulatory, you know, the company's pushing for it as a form of regulatory capture. I think it's worth noting here that the AI risks that certainly the companies and I'd say a lot of the significant portion of the AI policy community is worried about and is arguing for regulation on are risks that come when you're building models that cost hundreds of millions of dollars, if not maybe soon billions of dollars to build and to train.

The way that the Biden administration started this, and I think it's a good sort of general rubric for it, is by doing this based on sort of compute thresholds. The amount of compute that goes into training a model corresponds to the model's, roughly corresponds to the model's overall capabilities, and it, you set a threshold that's pretty high. For example, about around \$100 million worth of computing power to go into a model, you're capturing just the frontier companies who are, you know, have extraordinary amounts of money rolling in and rolling out at all times. And you can allow, you know, 99% of the AI ecosystem to continue to push, to innovate, to develop, while capturing just this small band of actors that might be developing models that might carry certain outsized risks.

So I think that's, when we're thinking about regulation, over-regulation, it's really worth dividing out, say regulation that would affect these frontier risks by a very small number of

very well-financed actors from regulations that would build a system that is imposed upon every AI startup that's trying to build a clever application that could, you know, make people's lives better.

MACCARTHY: Yeah, targeting the regulation where the risk is obviously makes some sense, and that doesn't affect the entire ecosystem. But even among the, the top tier, there is a kind of prisoner's dilemma game going on, isn't there? And that, that's what was behind Dario Amodei's call earlier in September, where he said we need a coordinated pause in AI development, which could turn into a kind of regulate the development process for safety. Not just slow the process, but do it more safely. But the, the prisoner's dilemma problem is that if one company steps up and does that, then the other company rushes on ahead, and the danger is that other company will achieve a permanent and decisive advantage. And so, both companies have to rush on ahead despite the risks. That's why it had to be coordinated, and that's why you needed some sort of government controls rather than just, you know, "Trust me, I'll do the right thing."

Kyle any, any thoughts or reactions to that prisoner's dilemma problem and whether the U.S. is likely to step up to try to address it in some fashion?

CHAN: Yeah. So one interesting thing about this prisoner's dilemma is that it is a fractal applying both within the US, within China, and also at the international level between the US and China. And so you have this really gnarly coordination problem then if you are concerned with out of control AI development.

And I think a couple things that are interesting to point out here. One is that as I mentioned earlier, the existential risk fear is not so great in China. And in particular, the fear of RSI, recursive self-improvement running away and sort of getting out of control where AI systems improve themselves and can do so at an ever-accelerating rate to the point where humans are increasingly marginalized in the process and maybe automated away entirely, and then you get AI systems, you know, you get this super intelligence explosion. This is the fear of a lot of folks who are concerned about existential risk. And I think that particular loop is not really a major sort of near term risk in China or seen as such. There are certainly labs who talk very explicitly about doing RSI and trying to close that loop, but I don't think that they see it as imminent as those in the American labs do and in the American AI community do. So that's just one starting point. And because of that difference, I do think that calls for a joint slowdown between the US and China will kind of miss this gap in perceptions of at least the RSI existential risk.

At the same time though, I think that we're now entering a point in the technology where it's not just sort of a race to get ever better capabilities, but we've kind of crossed a threshold where now these models and these AI systems, and especially with a good harness, can do real damage potentially. And so, you know, when you reach that threshold, then the question becomes like what do you gain as a nation, as a society from, from allowing that further escalation and what do you start to lose in terms of getting more risk from that.

I understand from the company's standpoint they might still be kind of caught in this in this prisoner's dilemma. But from a national regulator standpoint or from sort of a, a national policymaking standpoint, you know, it's this trade-off between getting ahead nationally versus getting a handle on the technology and not having blowback on your own companies or your own society. That is something that needs to be sort of balanced and figured out.

So I think in many ways both countries, both the US and China, are starting to see a shift in maybe some of those perceptions at the at the national level. And then ultimately it will be the national regulators themselves that will have to be the ones to coordinate this. I don't think the companies alone, and I think they're right. The companies alone can't do this, and their incentives are not, are not aligned to, to do this on their own.

LEWIS: And Mark, you've talked about, you've talked about the, the prisoner's dilemma. You might wanna talk about the Formula One race dilemma. Because when the yellow flag goes out, no one is allowed to pass the leaders, and that's certainly the prism at which some people look at the proposals from the leading companies. This is a way to reduce competitive risk from new entries to the market. So yellow flag sounds good for the Chinese they say it is, this means the US wants to stay in the lead and we're gonna be in inferior position. Not a good way to sell stuff.

MACCARTHY: That's not a bad analogy. You know, if you need someone to say here's the point at which you all will will not be allowed to jump ahead in the race."

Matt, I mean, where are we now, Matt, on, on the, on AI safety cooperation? I mean, we know roughly what the, the stage setting is. There's supposed to be a meeting sometime in November. The, then we don't know who's gonna be involved in the meeting. We don't know who, what the issues are gonna be. What do you think is going to happen in these discussions? Will it be good technical talks? Will it be at the political level? Maybe we think about threats from non-state actors as the, you know, something that we could both worry about and focus on. Matt, what's your take?

SHEEHAN: Yeah, there's sort of a version of what I think would be an ideal and, not just ideal, but, like, a feasible approach to this, one that takes into account the deep suspicion, enmity, and, like, very real constraints on action. And then there's what I think maybe we're likely to get, at least at this stage. When I'm talking about this or, you know, writing memos on this, I don't even use the word US-China cooperation on AI safety because I think that really is in some ways overselling what's possible. I think given the mutual suspicion, the best we can hope for, at least in the immediate term, is something that is short of cooperation and more than sort of this, how to put it, like unbridled competition.

It's not so much that we're trying to restrain, do mutual restraint on actions, but that we're trying to have technical exchanges that can inform the safety practices within each country without turning it into something that's binding between the two, without it being totally reciprocal. We're talking about, you know, sharing best practices on how do you evaluate for loss of control? How do you actually build sandboxes that are actually going to constrain these models? How do we deal with the fact that now the models are aware when they're being evaluated and thus behave differently in evaluation than they do in the real world?

These are, like, very open technical questions that both sides need to sort of get up to speed on, need to formulate their own answers to. And it's possible to have, I think, detailed and useful conversations about these without it being a necessarily a cooperative in the sense of, like, working together or mutual binding reciprocal restraints. So I guess that's just, that's my sort of model of how this could be done.

I think the main curveballs, you know, wild cards here are the political ones. If AI safety is deemed to be a hoax by the White House and that becomes sort of a blanket policy, then

honestly, there's probably not all that much to talk about with China on this front. I think the overwhelming majority of AI issues are going to-- the technology is going to be built in each country, the problems are going to be confronted in each country. We don't actually need to be talking to each other about too many of these things. It is the more frontier risks, cross-border risks where there is something of a need for this. So that political curveball is, I think, maybe the biggest factor that we don't know.

If I were to just make a some type of a prediction about what might happen in November is, one, I think the sides now have a little bit more time to prepare for this. In the run-up to the September one, despite the fact that there was this time between May and September to theoretically get ready for this, we were just-- the, the two sides were just not getting ready for it in very, like, practical senses. There, there was not a lot of preparation done. So maybe this sort of kicking the can to November is actually a good thing. It gives time to, to have the preliminary meetings.

And then the two deliverables that I think would be good and meaningful would be, one, turn this into a structured ongoing dialogue. As you mentioned, all the pieces on this chessboard are moving. The technology is moving. How each side sees the risks is moving. And my prediction is that reality, the reality of the technology is going to force some updates on both sides. And as those updates are happening, you want to have a structured, well-run dialogue where we can sort of digest those and communicate about those in tandem. So turn it into something structured, ongoing, sort of fully staffed.

And two, potentially doing something around this, I wouldn't say necessarily a crisis communication hotline as it's a bit, a long history of failure basically, but some way for the two sides to communicate about evolving, rapidly evolving AI incidents that are crossing borders or potentially crossing borders. I think those are the two things that I would hope from this first November meeting, and I think are totally feasible.

MACCARTHY: Yeah, that, that sounds great. Matt you've used the phrase before, "AI safety in parallel" to describe what's going on rather than some sort of negotiated deal. But the addition is that you have a nice communications channel and a sharing of information. It's not just I do my thing and you do your thing, but we actually try to do something that communicates what we're doing to the other side.

Kyle, do you have any thoughts or reflections on these developments? Anything you wanna add to what Matt laid out?

CHAN: Yeah. Matt has a bunch of great points. I agree completely. I think one thing to just add is one thing that I think folks who don't follow China closely should just keep in mind is that it matters a lot to have the official dialogue, to have the buy-in from Beijing on the Chinese side.

I think the Chinese labs themselves are nervous. You know, I like to joke that they're afraid of the US government, and they're afraid of the Chinese government, and I can tell you which one they're afraid of more. And so they will be hesitant to do anything on their own to take the initiative. If they have some kind of incident, they will look to the government most likely about whether to share that information with the US side.

And so it's important to have, you know, even if it feels kind of thin at the beginning, it's important to have this official dialogue because I think from the Chinese side you know, we underestimate from the American side how much, you know, we kind of do things and companies do things and people do things sort of on their own initiative often, you know, ahead of the government and that's kind of a, a different way of dealing with a lot of these issues. But on the Chinese side, there's a lot of reluctance to do that. So it matters to have that kind of green light from the top in, in China to say, "Okay, this is something that we want to engage in. And now, let's start talking."

LEWIS: When we had discussions of a hotline on cybersecurity, this was a few years ago, with the relevant Chinese agencies their answer was-- first of all, we had a hotline and they never answered. So that was number one. Number two, they said okay, we'll answer if you give us two weeks notice through diplomatic channels." Right? So I think their preferred venue is diplomatic channels, and we'd have to persuade the Chinese that there's some urgency that requires us to go beyond that channel for any sort of communication.

MACCARTHY: I heard something recently where someone said, "If you don't like a hotline, why not just send them a fax? That way they can you know, absorb the information, think about it and respond in, in that way."

We've got some questions from from the audience that I wanna turn to now in the last few minutes that we have together.

And one, one question that came in is -- and this may be for Matt and for Kyle. What independently verifiable safety evidence could the US and China agree to exchange realistically without sharing all the model weights and the sensitive information and so on? And of course, that has to be updated as the models themselves change. What really could we share between the two countries without giving away the store?

SHEEHAN: Yeah, I think the independent verification part of that is the hardest part of that. Verification can mean a lot of different things, you know, within AI and then obviously across fields. And some of the more ambitious, maybe potentially fantastical proposals right now around a sort of a coordinated global pause on AI, they often revolve you know, they revolve around this idea that we are going to be able to verify whether or not sort of large amounts of compute are used for training models or doing inference, sort of deploying the models. And, you know, we just find some way to verify that we're all training, not-- we're all doing inference, not training.

That technology is just not there currently. There are startups, there are labs that are working on this, but I don't think even the most, you know, ambitious of those would say, like, we're ready for verification right now, and certainly not of the quite detailed issues around what safety testing was done and what were the results. I think you can share that sort of in good faith if that, such, such a thing exists.

You can share information about what you did, and by looking at the other sort of the sophistication of the information that the other side shares, you can learn something, I think. Oftentimes in our dialogues, you know, we'll hear maybe the, the Chinese side say no, we really care about X issue." But if you kind of probe a little bit more, you're like, "Eh, it doesn't really seem like you've really thought this, thought very deeply about this." So I think there is information you can glean from these type of exchanges, but the independent verification

part of it, is not there currently and might take more time than, than people think that we have.

MACCARTHY: And Kyle, any thoughts?

CHAN: Yeah. I would-- I'll build on what Matt said. I think the term I think of is selective sharing. I don't know if we'll really come to some binding agreement about above what threshold, how much information, exactly what sort of technical specifications we're going to disclose with each other. But at the very least, we can begin to share a few selective bits of information about, say, incidents of concern. And so, you know, there could even be like a, a joint review of the Hugging Face incident based on publicly available information, right? So you don't even need to dig into sensitive you know, information here. We can start with that.

And from the Chinese side, you know, a lot of folks in the US I've talked to in Washington are like, "Well, what if China doesn't share anything?" And my baseline expectation is they might not wanna share very much, especially at the very beginning. But starting to build that muscle, where we share with them some of our concerns, get their reaction to that, see if they have any way of addressing these. Maybe we can encourage them to start to share incidents of their own. And they might share something very low-level, where they don't feel like they're risking their reputation or other sort of political risks from their standpoint. And then we might get into more of these cases where we can get into sort of more meaty incidents and, you know, build up from there.

But I think, you know, we don't need to start with a perfect arrangement from the beginning. We can start with kind of this piecemeal solution and build this as we go along.

SHEEHAN: Can I just add one thing very briefly? I agree on the incident sharing, and I should have mentioned in my answer like testing methodologies, testing and evaluation methodologies. We do not wanna teach China how to build better models, but we do want China to be able to know whether or not their models pose a risk of loss of control or pose a risk of sort of uplift in the chem bioweapons field. Like we, we want them to be able to figure that out for themselves because that will update their own risk perceptions, and that will make it more likely that they actually address these risks that can cross borders.

So I think testing methodologies, I spent a long time talking to different people in the testing evaluation field around, you know, when you share information on testing, can that indirectly enhance the capabilities of the model? There is a certain amount of gray area there, but the conclusion that I think is sort of spreading across the field to a certain extent is it is possible to share a certain amount of testing information that doesn't enhance capabilities but does enhance the other side, China's ability to, their own risk awareness and ability to potentially mitigate those.

MACCARTHY: Yeah, I think the-- Matt, you've mentioned in a number of circumstances that China's regulatory system seems to be iterative. They put out a proposal and then they adjust it in light of developments. That may be what's gonna happen in the cooperative arrangement as well. You start off with something and then you have to adjust it as time goes on.

I've got a question from the audience that raises a very general question and maybe Jim you could jump in on this. A lot of people say that the right framework for thinking about this is arms control. We need to have an arms control-like agreement to control AI risks. Is that a suitable framework for AI governance between us and China?

LEWIS: The short answer is no, right? There're a couple reasons for that. The first is the Chinese don't have that long experience with arms control that we have with the Russians. And so when you negotiate with them on arms control issues, their attitudes towards risk are very different. Even regarding nuclear weapons, but certainly on other WMD. Very different approach.

Second, we don't have the data for arms control negotiations. So I've said this a couple times. Maybe this new NVIDIA system will build the data we need, but you don't have you don't have the... it's hard to negotiate when there's no there there.

And finally, the approach that's been put forward as an alternative is maybe a reliance more on standards, on principles, on guidelines. That's something there are international bodies that do standards now. We could talk about AI safety standards in a way that the Chinese would want to participate. They aren't gonna wanna do arms control. They have very mixed views on negotiating with the US at a serious level on security issues. And AI, there's not enough there there. There's not enough meat on the bone to justify them taking an additional step.

MACCARTHY: Comments from the, from the, you others? Matt, Kyle?

SHEEHAN: I'll just maybe emphasize again the, like, the verification component of this. It's, if you cannot verify a sort of mutually binding agreement, then the, the default assumption is going to be that the other side might be violating it. And so in some ways if that verification technology is maybe two years out optimistically, maybe further out than that, then we might need something sort of in the interim.

And yeah, in some ways I'm a little torn between sort of, you know, the ambitious proposals that are essentially planning for a world in which these risks become very real, very serious, and essentially impose themselves on the two countries. I'm torn between sort of the, you know, that version of contingency planning and a version of things where we essentially work with the current risk perceptions, we work with the current state of play, in which case, yeah, arms control is the wrong frame. Arms control would only become the right frame in a world where the technology, the risks, and our understanding of those risks dramatically changes, and then maybe where we have verification capabilities on top.

LEWIS: And maybe building on that point really quickly, don't involve the UN. Adding the UN to this mix is not gonna lead you to any sort of progress because you'll have two players and 191 strap-hangers.

MACCARTHY: Maybe that's at the end of the process, but certainly not at this point.

I have one question from a press person that I wanted to sketch for you guys, and then maybe you can combine your answer with your whatever final comments you wanna make. And it goes back to the distinction we had at the beginning of the program between China seeming to be more focused on applications, market-specific items, and so on, and the US

seeming to focus much more on a general purpose technology you know, aiming for excellence in a broad range of tasks rather than just looking at the applications.

And the question is, if we accept that what does that mean in terms of how the two countries are gonna regulate their markets, and what might it have to do with our ability to work together on AI safety cooperation? Kyle, you want to jump in and take a shot at that?

CHAN: Sure. Yeah. I think there's many different levels that policy can operate in the, in both countries' ecosystems. And so it'll be interesting to see whether, for example, in China already some of the AI regulations are geared towards services more than models. And so I could see that kind of framework already being adapted for AI safety for a range of applications.

At the same time though, you know, we may end up having to reach deeper into the pipeline. I think in the US, for example, there's a lot of focus on the new model release as being that crucial threshold. And increasingly we need to think about even before the models get released to the public these recent incidents or a whole bunch of incidents we see are not actually public models that have caused issues. It's about testing these models. And so, yeah, it'll be interesting to see whether there will need to be some rules or some efforts to regulate the development and testing of the models themselves within the labs.

And then there on the Chinese side, it'll be interesting to see. I have no doubt, and I think many in China have no doubt that if Beijing wanted to they could intervene quite heavily into those AI companies internally. The question is, would they be motivated to? What reasons, what risks would kind of spur that action? And then what, what form that would ultimately take?

MACCARTHY: Got it. Matt, any final words from you?

SHEEHAN: Yeah, I think Kyle's suggestion of like in the US, at least in I'd say maybe the more national federal level discussions on AI policy and regulation, the focus tends to be on the model. And in China, the focus is on applications, and it's usually on service provisions. So their generative AI regulation explicitly excludes any generative AI thing that is not facing to the public, that is not a person interacting with a chatbot. Companies using them, internal deployments within the AI companies, those are all excluded. They're focused on applications. They're focused on sort of service provision.

That could change. I think this is another one where it's, this is all going to be very responsive to reality to a certain extent. And maybe we see that and we see that the architecture, the regulatory architecture that China has built has then sort of turned towards these other questions and they kind of have a running start. Or maybe they're slower on that front and our focus on the model has sort of set us up well for moments like that.

MACCARTHY: Maybe one area where we could make progress, as Kyle suggested, is to turn the attention to the risks that arise at the at the development stage before the models are released to the public. That might be an area where we both could benefit from some conversation. Jim, you wanna wrap us up and some final words?

LEWIS: So I think I'm gonna say something probably never before heard at a Brookings seminar. In this case, I think the president's right, right? This is the US' lead to lose, and so

we need to look at it from that perspective. It's not like arms control or the Cold War because the opponent there was a peer. I'm not sure, and the others have sort of suggested this, that China is a peer yet in AI. They're very good, but we certainly dominate the markets by about two-thirds. So that adds a different perspective to this.

I don't like bans. Bans were discussed routinely during the nuclear era, and they were always rejected because a unilateral ban gives the advantage to the other side. That's pretty clear too. So when we assemble more data, we can start having serious discussions. The dialogue in November is a useful starting point, but we shouldn't have too many expectations of reaching some kind of agreement.

MACCARTHY: So Jim, thanks for ending us on a controversial note. We'll have to leave it there. And thank all of you for your comments and sharing your insights and expertise with us. And thank you to Brookings for putting this event on, and thank you to the audience for listening. We'll see you next time for further discussions of this interesting and complex set of issues.