The Brookings Institution
***The Current* podcast**

"Can we slow down AI without losing to China?"
September 21, 2026

Participants:

Thomas Wright
Senior Fellow, Foreign Policy, Strobe Talbott Center for Security, Strategy, and Technology
The Brookings Institution

Elham Tabassi
Director, Artificial Intelligence and Emerging Technology Initiative
Senior Fellow, Global Economy and Development
The Brookings Institution

Kyle Chan
Fellow, Foreign Policy, John L. Thornton China Center
The Brookings Institution

Episode Summary:

As AI models grow more advanced, warnings of catastrophic risks are escalating. But are existential fears overshadowing immediate threats? Brookings experts Elham Tabassi and Kyle Chan join Thomas Wright to evaluate the reality of runaway AI, the debate over government regulation, and the geopolitical stakes of the U.S.-China technology race.

WRIGHT: How do we get from the systems that we have today to something genuinely catastrophic?

CHAN: The concerns about existential AI risk are premised on the idea that we will hit this ever-accelerating feedback loop in the very near future, something called recursive self-improvement, where AI systems improve themselves and they do so faster and faster to the point where humans no longer have control over this process.

[music]

WRIGHT: You're listening to the Brookings Current, part of the Brookings Podcast Network. I'm your host, Tom Wright, senior fellow in the Strobe Talbott Center for Security Strategy and Technology. Something unusual is happening in the world of artificial intelligence. Over the last couple of weeks, some of the people building the world's most advanced models, Dario Amodei, Sam Altman, Elon Musk, and Demis Hassabis, have all warned that the race to develop ever more capable models may be moving too fast and that the industry needs to slow down.

These warnings come at a time when the U.S. intelligence community describes AI as a defining technology with growing implications for cyber operations, defense, and warfare, and they stress the importance of keeping humans in control. And then there's the extraordinary Hugging Face incident. During cybersecurity testing, OpenAI says that its own experimental AI agents circumvented safeguards, reached the internet, and compromised systems at OpenAI and Hugging Face.

New reporting suggests that some of that activity began earlier than previously known. Other incidences have also come to light in recent days. Soon, President Trump and President Xi will meet at the White House, and AI is on the agenda. Can we slow down AI without losing to China? That is one of the questions that we'll be looking at today, and I am delighted to be joined to unpack these high-stake questions by two of my Brookings colleagues.

Elham Tabassi is the director of Artificial Intelligence and the Emerging Technology Initiative here at Brookings, where she leads our research on AI safety, standards, and governance. And also with us is Kyle Chan, a fellow in the John L. Thornton China Center, and his work focuses on China's tech sector and industrial policy.

Elham, Kyle, welcome to the Brookings Current.

TABASSI: Thanks for having us.

CHAN: Great to be here.

WRIGHT: So let's start maybe with just an overall assessment of where we are. Are the AI executives right to have expressed the concerns that they did in recent weeks, and what do we make of it? And when people say the latest technological developments have changed their assessment of AI risks, what exactly has changed?

So Elham, maybe we could start with you.

TABASSI: I think we should take them seriously. They are building this technology. They know what's happening inside their companies better than anybody else. So I think at the minimum, you should definitely take them seriously. And the second part of your question, I don't think it is any of them by itself, but the combination of all of them that these AI algorithms, these AI models are increasingly become more capable.

They are better in coding, and because of that, they are better in exploiting weaknesses in the codes. They can coordinate, the agents can coordinate among themselves. So all of them together is getting us this sort of capability and the things that we hadn't seen before, and because we hadn't seen before makes it more difficult for us to figure out how to measure the type of risk and impact they can have.

WRIGHT: Great. Thank you. Kyle?

CHAN: I think that in some ways we need to focus on some of the more immediate risks. The concerns about existential AI risk, the fears that AI might kill us all in the next decade, are premised on the idea that we will hit this ever-accelerating feedback loop in the very near future something called recursive self-improvement, where AI systems improve themselves and they do so faster and faster to the point where humans no longer have control over this process.

There are a lot of steps there that remain to be figured out. This depends not just on better software, but on better data, better compute infrastructure. There are a whole bunch of missing pieces along the way. So in terms of the existential AI risk, this is something we should be thinking about now, but I don't know if that's our top priority.

There's a whole bunch of other risks related to how AI systems today can attack other companies, can perform advanced cyber operations, can be used by other nations or non-state actors to carry out, say, ransomware attacks, or perhaps even develop certain kinds of biological pathogens. These are risks that I think are rising to the fore, have been talked about for a long time in the AI safety community in the U.S., but are now being taken more seriously by some of the policymaking world.

I mean, the risks you identify there are largely by actors, by human actors, using or abusing, you know, AI models to conduct, you know, cyberattacks. But the news over the last few weeks, particularly on the Hugging Face incident, was of autonomous sort of agents. So some people would argue we're already, you know, in a world of recursive self-improvement.

You know, we're already seeing, these models act sort of autonomously in ways that we don't fully understand, and it's only going to accelerate from here. So do you disagree with that?

I would separate this out into agent autonomy versus the recursive self-improvement. The fear from recursive self-improvement is not just that these systems might go about doing things that we didn't intend for them to do, but that they can actually improve themselves faster and faster, become smarter and smarter, and augment their own capabilities at a pace that starts to spin out of control.

And I think that is ultimately sort of that for some people's view, that near horizon risk that could really wipe out humanity, if to put it bluntly. Versus some of this autonomous action, the sort of current degree of runaway AI systems doing things, hacking companies that they're not supposed to do.

That is a problem today, but I don't think it's seen as a existential risk. There is not a risk today that those systems will say, go about carrying some broader takeover of the entire grid or of military weapon systems. So I would separate some of those concepts out.

WRIGHT: Okay. Elham, if we could turn to you just on this existential question, how do we get from the systems that we have today to something genuinely catastrophic?

Like, what precisely happens in that scenario?

TABASSI: We have evidence today that the current AI systems can behave unpredictably, can exploit weaknesses in safeguards, and can show capabilities and behaviors that is hard to predict. But none of this demonstrate that AI systems can, as Kyle said, you know, run away, the loss of control scenarios and those extreme existential risk, catastrophic risk that has been talking about.

In terms of the recursive self-improvement, also agree that the recursive self-improvement right now is bounded. We still have the humans to set the objectives control resources, evaluate the result, and the question becomes can that loop get closed and accelerate? And I think that, as Kyle mentioned, depends on data, compute, energy, and those are bounded by physical and economic timelines rather than software timelines.

The way the community are talking about is sort of two ends of the spectrum. One is, as you talked about, a bad actor sort of abuse, misuse AI and try to launch a cyber or bio attack, as Kyle explained. You know, companies are doing a lot of things to prevent this type of abuse and misuse of AI systems.

But AI systems can be very smart and very dumb at the same time, and eventually they can jailbreak and all of those things can happen. The other end of the spectrum, as some in the community argue, is AI go run away, go rogue and do things and the example that is they bring is that they have a objective and in their achieving their objectives, they need resources.

They start competing with the human on the resources and strange, bizarre, crazy thing that can happen. And there is a whole sets of scenarios that can happen in between that it is not just one thing, but a gradual and combination of the things that can happen. So we see that, for example, some of the things that they talk about that can happen is that, well, we see that AI is being increasingly deployed to do the research and engineering.

Then we have the sort of the giving more autonomy that Kyle also talk about this, and it just can be because the companies are under resources, and it's easier to give more autonomy to the agents to do that work. Then on top of that, we have the issue

of the alignment and, you know, is the algorithms, is this model are exactly doing what the humans set the objective for them to do?

And as they become more capable, the, the alignment problem becomes more difficult. And then we have the evaluation failure, and as you alluded to that, more and more have situational awareness, so they know they are in the testing environment versus the real world deployment, so it becomes AI evaluations much, much, much harder. We as a society are increasing our reliance on AI more and more, so AI becomes integrated into the critical infrastructure where continual monitoring or even shutdown if something happens becomes harder.

So a lot of the things in the scenario that I mentioned can be assumptions. Some of them are measurable, some of them are active area of the research, but if you put all of them together, that something bad and catastrophic can happen. But what I want to end with is the importance of the measurement, the importance of accumulating evidence for all of these things.

And we would be much in a better place if we can have a better measurement systems, know the indicators, be able to measure all of these warning signs, and be able to intervene as they come to prevent all of them.

WRIGHT: Yeah, just picking up on the point you raise that the advanced models sometimes know they're being evaluated and change their behavior and might act differently when they are not being evaluated.

I mean, what is the answer to that? Is there a way to sort of assess these models to make sort of judgments on safety, or are we getting to the point where that's not technically possible?

TABASSI: It's a challenge. It's something that we haven't been, you know, in testing of the AI systems and testing of the softwares, we hadn't seen it before. We definitely need better, more rigorous evaluations to figure out how to do that and when they start changing behavior, but continuing the testing and continuing the monitoring of the systems while they are in use to detect all of these failures or flaws sooner.

WRIGHT: Yeah. So question to both of you, maybe starting with Kyle. I mean, just building on all of that, I mean, do you think that there ought to be, you know, government regulation or, you know, a license required to issue new frontier models given all of these risks?

CHAN: In a word, yes. We need some kind of regulation, and up until this point, a lot of this has been left to the industry, to the private sector to figure this out on their own.

And to their credit, because they are genuinely concerned about these issues, they have developed their own internal safety protocols and efforts to do the sort of measurement and testing and to develop guardrails. But this is not enough. I think incidents like the Hugging Face incident show that the industry needs to do more to get a handle on some of these risks, and that ultimately, these should not be voluntary industry limits, but something that the federal government needs to be involved in, and perhaps even at an international scale.

WRIGHT: Elham, do you agree with that?

TABASSI: There is a lot from other industries to be learned right now. I completely agree that there is limits to any sort of the voluntary solution. I still think that the voluntary solutions and voluntary mechanisms play a really big role, and really here talking about the standards, because standards can also be voluntary because, as we know, the systems is changing too quickly, so we have to be able to adapt.

So we need regulations on measurable risks rather than a hypothetical label, and it's also important to do it in a way that it can be adaptable to the rapid changes that's happening in this environment and policies that stops boxing us into the solutions that cannot be.

WRIGHT: So David Sacks, who until recently was the AI czar at the White House and now still advising the administration on AI and technology, has sort of made the argument that government regulation requiring a license is just basically a means for a set of people in the AI community--effective altruism--to control AI politically and ideologically.

And it seems the president of the United States agrees with him. There should be no government regulation. It should be-- We should just rely on litigation, right? So if a company releases something that's not safe, then, you know, they're liable to be sued, and they will get in legal trouble and be punished maybe by the market as well.

And what is your answer to that sort of concern that, you know, who is regulating? How would you address or would you address the concerns of those, particularly on the right, who would worry that it's basically their political opponents that will be controlling this new technology?

TABASSI: So the word licensing implies there means the threshold for capability or risk that needs to be licensed, and then also a sort of the verification to see that if the thresholds is met, and also evaluation and measurement mechanisms for enforcing the regulations.

So for a very long time, I always have been saying that, while I agree with the regulations as sort of a backstop we also want the good regulations. Smart regulations are those that can also be enforceable. And currently, we don't have that measurement infrastructure, that evaluations, verification, validation mechanisms to be able to do the enforcement of-

WRIGHT: Yeah, but who would decide all of this?

TABASSI: Right. So who would decide all of this? And I would go back and say that there is a shared responsibility across all of the value chain. Companies that are building the AI systems, they know a lot about the systems, about what systems can do, and also how to test them. We need external independent evaluators to be able to verify those claims, and we do need the government to set the rules, to set those, those thresholds.

And it's very difficult because we just don't have the terminology and the definition that we need. You know, people talk about superintelligence, but actually what

superintelligence is. People talk about kill switch, but kill switch cannot help with the open weight. So we need representative and different minds more.

WRIGHT: Yeah. I guess one way of framing the question, if there was a licensing system and there was strict government regulation, but Donald Trump decided that the people who would be determining that were Elon Musk and David Sacks, I mean, how comfortable would we all be?

We'd probably be worried, but it's flipped on the other side, like they are worried, right? And they are basically in charge at the moment. They are worried about by setting this up, it will be people they don't agree with. And getting out of that trap, I think, is a tough one. But Kyle, why don't you address that quickly, and then we'll turn to China.

CHAN: The problem about regulatory capture is a real one for any industry, for any kind of regulation, and I think that also applies to AI. What's crucial to deal with that is to make sure that there are a diversity of views and voices that are feeding into how this regulation is created, that it's not dominated by a single company or even a single layer of the AI industry.

For example, the American AI ecosystem is much bigger than even just the two model companies, OpenAI and Anthropic alone. So thinking about different parts of the ecosystem-

WRIGHT: But how do you deter- how do you break the tie? Like, Zuckerberg's out there saying basically, "Go fast," right? No regulation Musk seems to be somewhere in the middle, can't quite tell.

And, you know, Altman and Amadei are saying, you know, "Slow down." You can have all of those people in the room, right? But then if you're president or you're Congress trying to determine the system, I mean, there has to be a tiebreaker, right?

CHAN: Well, I wouldn't see it as a tiebreaker because in the end, this is not for the industry to decide, and it's not a power struggle

between the industry that should decide a regulation. Ultimately, that should be decided by policymakers and voters. And so what we want to have is we want to have input from a wide array of sources, including within the industry, but also including outside of the industry, and have that feed into a broader policymaking process that we can discuss together publicly, because this is something that doesn't just affect any one company or any one industry, but can potentially affect all of us.

TABASSI: It cannot be just industry sets the rule for themselves. And AI systems are sociotechnical. It's just not a technical problem. It's not a technical problem with a governance footnote.

WRIGHT: Thank you. Let's turn to China. President Trump and President Xi are meeting soon. AI will be on the agenda. You know, we hear a lot about how both countries should cooperate on AI safety.

But in a New York Times essay today, Seth Center, who led those talks for the Biden administration and served in the State Department, wrote that in 2024, his

experience was that the Chinese side did not take AI safety seriously and only saw the opportunity to jockey for strategic advantage. And let me quote him here because I think it's important.

He said, "If both sides meet in earnest, the very best that I could envision is a non-binding commitment in which each side would agree to put pressure on its companies to publish and improve safety benchmarking for cybersecurity and biological risks, and to disclose rather than bury reports of concerning incidences."

So I guess, Kyle, to start with you, I mean, is he right in his pessimism, and if not, why not? And, I mean, how do you think China sees AI risk, I guess, at the community level, and do you think that is how Xi Jinping also sees it?

CHAN: So calls for a joint AI slowdown between the U.S. and China are not gonna happen.

China's not gonna agree to this, and frankly, the U.S., the Trump administration is not keen on this either. There is a long way to go if we're gonna talk about binding international treaties or agreements, a kind of arms control for AI. A lot of people make the comparison to nuclear weapons control. That's not the right framework in my view.

Instead, I think we have to take a step back and see this as a much longer road and a much larger set of issues than simply do we regulate or not? Do we slow down or not? Do we cooperate with China or not? And so I would start by thinking about how China sees AI risks and how these have been changing over time.

I think Seth Center's piece highlights a moment in time for Beijing where they saw AI as a huge opportunity and did not really emphasize so many of the risks. And so when the U.S. wanted to talk about some of these risks, I think the Chinese side saw it as a diplomatic issue rather than a national security issue.

That's changing now. In Beijing, there is growing concern about some of these larger scale AI risks. They're no longer just purely saying, "Hit the accelerator and go faster," although they are still trying to catch up to U.S. models. They're getting more concerned about cybersecurity. They are watching very closely incidents like Hugging Face.

They're worried about autonomous agents getting out of hand. Some of their recent regulations around AI agents are targeting autonomous systems and trying to tie back human responsibility to some of these. And when you look at what some of the top Chinese officials are saying domestically within China, for example, China's head of their Ministry for State Security recently laid out a whole bunch of what he saw as major AI risks to China's own national security and to the party's control on power, and those included things like cyberattacks that are enhanced with AI models.

He called out American AI models in particular as part of that. And he talked about a whole range of ideological and content-related issues that have long dominated China's approach to AI regulation. So I think what's happening now is we are certainly not at the point where China is worried about existential AI risk.

They don't think that AI will kill us all in the next decade. The policymakers don't think that way, and the AI industry for the most part does not talk that way. But I think increasingly they do see problems with these other AI risks that are no longer sort of the content social risks from AI, but threaten national security.

And on that front, I do wonder if there will be growing overlap between U.S. and Chinese interests in mitigating some of these, given that both countries care about this and neither country wants to see an American or a Chinese AI model, say, attack their critical infrastructure or be used to hack into a major hospital system.

TABASSI: I think Seth mentioned that in his article too, that the expectations of some sort of the agreement is, I don't think it's very realistic. But what we need is at the lower level, technical level, you know, we're talking about, you know, kind of mutual recognitions and mutual verifications, how technically we want to solve them.

So it would be better if at the technical level, the experts of the two countries or, you know, maybe even broader globally can agree on some mechanisms for the evaluations for verifications and validations and feed it up to the diplomacy conversations, and I don't see that's happening very vigorously right now.

WRIGHT: So let me just ask one more question on this. I mean, say Xi Jinping or the Chinese side basically say, "Look, we agree." As you laid it out, Kyle, that there are specific challenges that need to be addressed, but, you know, we can't really do that in the presence of these export controls on high-end chips.

And you say you're trying to keep us down and win the AI race, and that's, it's not practical for us to pace or to really get serious about a cooperation as long as that's the case, so we need you to lift those. Sebastian Mallaby has made some version of this argument recently. How do you think about that?

And on the flip side, you know, there are many people here who say, and Dario Amodei is one of them, who says that, you know, in order to buy time to be able to pace the frontier and to be able to assess safety risk, you know, we need to slow China down, deny them the compute power, not send over the chips, so that, you know, we're not in as much of a race, like we've a little bit more wriggle room, to actually slow down, you know, ourselves.

So where do you, starting with you, Kyle, maybe, but where do you come out on that question?

CHAN: I think we should keep the export controls on semiconductor chips to China as a separate issue to the extent that we can from these AI safety discussions. The Chinese side obviously wants to link them together.

There are some who have argued that we should use this as a carrot, and there are some who have argued in the U.S. that we should use this as a stick to pressure-

WRIGHT: What's the stick one? Sorry, I don't fully understand that. I mean, 'cause as I read it, they say basically just ought to have the controls for their own sake to slow China down.

You mean that it's leverage or?

CHAN: Yeah. The threat of cranking up the export controls could be used as a form of leverage to bring China to the table to talk about AI safety. I think that the right way to go about this is to keep the export controls as they are and as a separate issue, and to focus on trying to figure out areas of mutual overlap for the AI safety issue that affect both countries.

And it'll be tough. I mean, the fact of the matter is we're competing very fiercely to build better and better models, but also to use AI to augment a whole bunch of economic and military capabilities. So this is happening in China, this is happening in the U.S. I think that we're gonna have to figure out some way, like the U.S. and the Soviet Union once did, to not let the pure competition between the two AI superpowers be the only thing that we focus on.

We need to think about not just the risk to each other, but the broader risk from this technology to the rest of the world.

WRIGHT: Thank you. Elham?

TABASSI: From what I read and hear some of the wording in the Dario's letter hasn't land well with the technical and scientists in China about the things that can happen because any sort of the mutual recognition really needs a better relationship with.

WRIGHT: Yeah, I mean, I think also it wasn't just the essay or the letter, it was the national security report, the 134-page report that Anthropic released, I think last Friday, in which basically they show that Chinese actors were using Claude, you know, for their own purposes, but also they put in sensitive information, which then, of course, Anthropic got access to betraying Chinese national security secrets and all sorts of rumors flowing over the last week and on, on how the Beijing responded to that.

But we don't, I think, know exactly what happened yet. Let me finish with just a lightning round question, 'cause we've talked about a lot, but just what is the one specific and realistic policy that the U.S. should prioritize this year to reduce the risk of recursive self-improvement and advanced AI without actually killing American AI innovation?

So, Elham, maybe we could start with you and then go to Kyle.

TABASSI: Sure. I am a big fan of measurement and having evidence to support any sort of the claims, so I think we can just start by asking companies to report how much of their research is automated and how much of these recursive self-improvement, each cycle, how much gain has been achieved because scientifically, we talk about diminishing return.

So just have that data and get the evidence to build a rigorous basis for understanding the problem to guide

WRIGHT: policymaking.

Great. Kyle?

CHAN: I think we need to talk to China, and the most important thing at this stage is communication channels. We need what I call thick communication channels, formal and informal, between members of China's AI industry, AI safety community, and our own, as well as government to government discussions.

And the idea here is should an incident occur, should an OpenAI model not attack a company like Hugging Face, but accidentally attack a Chinese platform or vice versa, what would we do? How could we inform our Chinese counterparts that this was not intentional and exchange some views and vice versa? And so I think people who are aiming for some major international treaty are aiming too high.

I think we should see this as a potential ladder, and we have to start with the very first rung, and that's just talking.

WRIGHT: On the military side, we've been proposing for years, the U.S. has been proposing for years a sort of red phone, a hotline for exactly this type of thing, not on AI, but on more military activities.

And China's always rejected it, in part because they view those channels as sort of seat belts that encourage just reckless driving, right? And so, you know, our view was always that they had not experienced the Cuban Missile Crisis, so they haven't sort of fully appreciated the need for it.

[music]

Maybe that awareness will come with these risks. But Elham and Kyle, thank you both so much for joining me today to break this down. To read more of Elham and Kyle's research in AI governance, national security, and the U.S.-China technology relationship, please visit brookings.edu, and thank you for joining us on The Current.