Center for a
New American
Security

MARCH 5, 2021

TRANSCRIPT FROM CNAS TECHNOLOGY & NATIONAL SECURITY PROGRAM EVENT

CNAS Event Transcript: Artificial Intelligence and the Role of Confidence-Building Measures

CNAS Event Transcript: Artificial Intelligence and the Role of Confidence-Building Measures

I. Opening Remarks

Paul Scharre: Hello everyone and welcome. I'm Paul Scharre. I'm a Senior Fellow and Director of the Technology and National Security Program here at the Center for a New American Security. Thank you for joining us for today's discussion about artificial intelligence and the role of confidence-building measures. I'm joined today by a very exciting group of panelists—Helen Toner, who's the Director of Strategy at the Center for Security and Emerging Technology. Thank you, Helen. Welcome.

Helen Toner: Thanks, it's good to be here.

Paul Scharre: Kerstin Vignard, who is at the UN Institute for Disarmament Research, where she is the head of the UNIDIR Support Team for UN cyber negotiations. Kerstin, thanks for joining us. And Michael Horowitz, Richard Perry Professor and Director of Perry World House at the University of Pennsylvania.

Michael Horowitz: What's up Paul? Happy to be here.

Paul Scharre: Well thanks, thank you all for joining us for this discussion. A couple of interesting developments on how people think about AI risks and the potential for confidence-building measures. Most recently this week, a report released by the National Security Commission on AI—a monster report, very comprehensive—dedicates a whole chapter to looking at issues surrounding the risks of AI-enabled and autonomous weaponry.

Paul Scharre: I would also be remiss if I didn't mention, Mike and I earlier this year released a report on AI Risk and Confidence-Building Measures through CNAS that was funded by a grant from Carnegie Corporation of New York—we're very grateful for their support—where we explored some of these topics. With that, I'd like to dive right into a discussion about some of these concerns.

II. Panel Discussion

Paul Scharre: Helen, the first question comes to you. The National Security Commission on AI report that just came out this week talks about "mitigating strategic risks associated with AI-enabled weapon systems." What's your take on what some of these risks are of military competition of AI?

Helen Toner: Yeah, so the report as you say, is quite comprehensive, so folks can definitely go look there for a long list. I think of two big categories of types of risks when it comes to risks of military competition in AI, especially from a strategic stability point of view. One of those, which I think is pretty easy to grasp, is the risk of proliferation or the risks of AI-enabled weapons of mass destruction—WMD. So, the obvious example here would be, for example, autonomous drones.

Helen Toner: A really important dynamic I think, is going to be whether bigger countries that are spending more on military technology—are those countries going to be developing new technologies that are disproportionally going to help smaller actors and potentially contribute to proliferation? That's one sort of whole category of concerns. A different category of concerns I think, is related to the pressure that I think militaries face to be trying to move at machine speed, if they can, so obviously machines in principle can work and can sort of "think faster than humans."

Helen Toner: So, if there's a chance that your adversary is going to be using systems that can go faster than that, that can get inside your OODA loop as people say, there's going to be this sort of competitive pressure to try and keep up with that. I think that gets really difficult when you start to think about crisis stability, when you're thinking about the lack of escalation thresholds when autotomy is involved. It's even more concerning when you think about the systems that we have today and the way that they are really not sort of secure, robust, resilient, interpretable. So especially with questions of deploying the types of systems that we have access to today, I think those concerns are sharped even further.

Paul Scharre: Thanks, I appreciate that, that's a good overview of a lot of the risks that we might face as we see militaries begin to use more and more AI and autonomous technology. Kerstin, I know you've been very involved in discussions at the UN since 2014 on autonomous weapons. What's the current international environment right now for regulation on either autonomous weapons, more narrowly or more broadly, AI-enabled military technology?

Kerstin Vignard: Thanks, Paul. I think the short answer—the prospect for hard regulations at the UN or coming out of the multilateral level are slim. But that's across the board on all arms control disarmament issues, not just on autonomy or on AI. First, we're simply not in a geopolitical environment where new legally binding regulations are going to be agreed anytime soon. But secondly, I think it's under-appreciated—there's a process issue that COVID has really brought to the fore.

Kerstin Vignard: COVID has impacted our multilateral negotiations in a phenomenally bad way and that's true for I think every sector. But diplomacy and negotiations is, I think, some of the most resistant to adopting online technologies and being able to conduct their discussion in a virtual format. So, for over a year now, meetings and discussions at the UN have been postponed, they've been canceled. And while some work has transitioned to informal online consultations—for example, the two cyber negotiations that I support have been regularly meeting in virtual information meetings since last spring—I'm in New York this week because next week, we will have the final negotiation session of one of these negotiations here at the UN—a physical meeting. That's because there are some member states who insist that physical meetings to negotiate, to take formal decisions, and that's across the board—that's in the Security Council, that's in the General Assembly, and that's in processes like CCW.

Kerstin Vignard: So, physical negotiations are obviously incredibly hard when the majority of the negotiators can't travel. The CCW has, of course, and the discussion on lethal autonomous weapons, has suffered from this. Late in 2019, we saw the establishment of a mandate for 25 days of discussion on LAWS between 2020 and 2021. With COVID, there was single five-day week of hybrid meetings last September, which some states, true to their positions across other UN fora, refused to participate because it was a virtual—not fully in-person meeting—it was a hybrid meeting. So, the November meetings were canceled and the session that was scheduled for next week was also canceled.

Kerstin Vignard: So, there's potential dates coming up spring, summer, to kind of resume the discussions. But the difficulty of gaining consensus on holding, just holding, a hybrid meeting is really up in the air. So what does this mean? Lack of substantive progress, due to resistance of the positions of some states, but also a lack of progress due to the process related delays, may lead to increased momentum for a like-minded process to ban autonomous weapons systems being taken outside the UN. The question is really, were that to happen, what important questions would be left on the table unaddressed and what measures would fill the gap in the normative and CBM framework among the most mature AI states who would reject such a process? Thanks.

Paul Scharre: I appreciate that. That's a helpful reality check on some of the difficulties of the international diplomatic process. Certainly if states are having trouble agreeing on whether or not to hold a meeting, that's a useful cold splash of reality on the prospects for actually kind of really hard limits on technology. Mike, I want to turn to get your take on the NSCAI report, and then we'll go back to Helen and Kerstin for their analysis.

Paul Scharre: The NSCAI report dedicated a whole chapter to talk about some of these risks associated with military competition in AI and then a number of practical recommendations. I should say as a disclosure I guess, that you and I both were consultants to the NSCAI while they were in development. Although I will personally say I can't really take any credit for the literally awesome work that the group had done—the Commission and the Commission's staff—but I'd be interested, Mike, in your reaction to some of the recommendations that they put forward.

Michael Horowitz: Thanks, Paul. Yeah, I don't think either of us is fully unbiased, although frankly they wrote whatever they wanted at the end of the day after asking us what we thought. But I think that these recommendations, while they won't please people in some quarters, actually represent a responsible and realistic swing at trying to manage the risks associated to strategic stability from advances in AI. In particular, the focus on the notion of the U.S. becoming a leader in promoting limits on the integration of AI into critical nuclear systems, the suggestion for a strategic dialogue with Russia and China—I mean, talking on the one end is like kind of easy, but I think there's something important there.

Michael Horowitz: Obviously, given the report that we wrote that was published last month, I was very pleased to see the integration of some of the recommendations we laid out surrounding confidence-building measures. I'll only speak for myself here, I think I've been skeptical at times about the prospects for formal arms control in the AI arena—seeing AI as a general-purpose technology, it's not like we had arms control for electricity. But that doesn't mean we can't buy down risk in really important ways.

Michael Horowitz: I think we can buy down risk with transparency measures and standard setting, rules of the road to try to improve behavior. I think we can buy down risk in particular usage areas and thinking about something like, say, an incidence-at-sea agreement-style agreement for autonomous systems. I was really pleased to see that the National Security Commission on AI report picked up some of those ideas based on their own research and laid out, I think, what would be a very active American approach to trying to decrease the risks that AI poses to strategic stability.

Michael Horowitz: Again, not everybody is going to be happy with those. With that, the report acknowledges for example, the possibility of autonomous weapons systems—it doesn't advocate for a ban on them, as I think some of the NGO community would have hoped. But I think certainly, if one looks at the American negotiating position over the years on many of these issues, this in some ways is the furthest that an American government—they don't represent the American government, they're a commission appointed by Congress—but that senior leaders like Eric Schmidt and Bob Work, and then there's several people from the tech industry, people that have gone into the Biden administration—this is a really good swing at ways to decrease risk. So, I'm supportive of a lot of what's in that chapter on strategic stability in particular.

Paul Scharre: I want to get your take next, Helen, what were your thoughts on kind of the NSCAI's...both their analysis to some of these risks and then their recommendations?

Helen Toner: Yeah, it was super interesting to see the final report come out, and I agree with a lot of Mike's analysis. One thing I'll say for anyone who, like me, was initially put off by the 800-some page count, it turns out that a bunch of that is blueprints and appendixes. It's something like 200-some pages of the actual chapters and they have a nice online interphase that you can kind of scroll through. So, you know, I thought I was going to have to strategize and sit down with a whole weekend and figure out how to break it up. It's pretty approachable actually, I think they did a really nice job. But it is very long.

Helen Toner: One thing I've been struck by also in just some initial conversations with folks—some folks from the NGO community as Mike mentioned, a few folks from Europe this week about their reactions to the report—is the interesting tensions between the different chapters, which I think reflect real tensions that the national security community in the U.S. is facing. So, for example, Mike, you just talked about the ways that it lays out these strategic risks, there's this whole chapter—Chapter Seven—about testing and evaluation, validation and verification, and the difficulties there.

Helen Toner: So, I was interested, given the existence of those chapters, that many of the conversations I've had with folks this week have focused more on things like Chapter Two, which is really focusing on pushing forward the AI-readiness of the Department of Defense at maximum speed and also some of the judgements that opened Chapter Four, which are really sort of standing up for the U.S.'s ability to build and deploy autonomous systems in battlefield contexts at some point in the future.

Helen Toner: Basically, I had these conversations with folks that were really saying, "Wow! This is an extremely pro-autonomous weapons report, DoD is really going full steam ahead, like gosh." So, I thought that was really interesting. Mike clearly disagrees with that take. But I thought it was interesting that depending on which chapters you read and how you interpret them, it can come across differently. I think this also, Paul, something I've heard you say plenty of times, is talking about the two big types of risks you can face when deploying new technologies—to go too slow or to go too fast.

Helen Toner: So, I think these tensions are reflecting that as well, that the National Security Commission in this report and, of course, DoD more generally, want to be saying to competitors and adversaries we're going to be going at world pace, we're going to maintain our superiority, we're not going to go too slow. While at the same time, needing to make sure that they are pacing themselves adequately. I think a couple of dimensions that I found really interesting here are the distinction between laying the digital foundations to be able to use these systems and build these systems, which I think is really the focus of Chapter Two. So that's more about data management, it's about human capital and training, and the digital systems within the Department, which are notoriously somewhat behind the times one might say, versus putting things into operational use on the battlefield which is just a very different question.

Helen Toner: Then secondly, also distinguishing between these more sort of enterprise potential applications, which, when we're talking about—again Chapter Two of the report—talks about bringing in more private sector products and things like that. I think if we're talking HR systems—we're talking these back-office functions—that makes a ton of sense, versus anything that's going to be really defense-specific—might involve killing people—really wanting to have far more caution and far more deliberate process involved there. So, I don't know that was some of the interesting thoughts that the report prompted for me.

Paul Scharre: That's great, that's super helpful. It's always good to know how people, particularly if there are folks coming from outside the defense community, perceive these kinds of things, because obviously these kinds of documents have multiple audiences right? They're audiences inside the government, the defense community that they're trying to message to and in many cases to spur people to action. There are audiences on the Hill, there are audiences in the broader tech sector, and then, of course, people internationally. I thought your reaction that you were conveying, I guess, others had to the autonomous weapons is really interesting in particular because when I looked at it, there's nothing from my perspective that's new or different right?

Helen Toner: I agree, yeah, I was surprised I sort of looked at it and thought yeah, of course, these are the things they were going to say. But it was a pretty strong reaction that I heard from multiple people this week, so I thought that was interesting.

Michael Horowitz: I mean not to speculate too much, but my bet would be that may reflect communities that just don't believe the U.S., when the U.S. says things like, "Hey, we're not developing autonomous weapons right now."

Helen Toner: I think there's also a really interesting distinction between something that I think Eric Schmidt, the Chair of the Commission, has said in media interviews this week and that is also in the report—which is, existing U.S. procedures are adequate to ensure that the U.S. will only field systems that are compliant with international humanitarian law. That is one statement—it's pretty easy to interpret that statement as saying we are on the way to doing this, we know how to do this, we have systems that are going to be rolling out soon—which it's not quite what he's actually saying. What he's saying instead is, there are existing checks in the process that are going to stop systems if they are not capable of doing that. I think that statement can be seen as, we're going full speed ahead, we have these systems in the pipeline they're coming through, our processes are great, don't worry we got this, which I think doesn't always land well for folks who are concerned about the impacts of these technologies.

Kerstin Vignard: I was really reading it kind of from an international perspective and talking to people outside of the United States. I think a lot of people were struck by the tone that it took, particularly vis-à-vis Russia and China comes out very strong in Chapter Four, obviously. I was really personally struck by the tensions that Helen described. I think a lot of people won't read the whole document; they'll dip into the chapter they're most interested in.

Kerstin Vignard: What makes the Commission's work so interesting and complex is those tensions that we see, between kind of those, in some ways, competing interests, even within the same government. So, I think we do ourselves a bit of a disservice if we only dip into Chapter Four and talk only about the autonomy discussion and where it puts us. So, I was struck by those tensions as well. But I have to say I really appreciated in Chapter Four specific, that the emphasis that the Commissioners put on looking ahead to what could be concrete CBMs. Mike, you mentioned some of this, for example, the practical risk mitigation measures, the automated tripwires.

Kerstin Vignard: The report specifically mentioned nuclear weapons but, of course, I think there's some maritime systems that would be really good candidates for this in the short-term as well. CBMs come in a variety of flavors and intensities, from transparency measures, to corporative measures, to stability measures. With the publication of 3000.09, in what, 2012? The U.S. took the lead in these discussions on transparency, right?

Kerstin Vignard: And being the first meant that the U.S. received a lot of criticism internationally for that directive. But putting it out there and in the public domain at the international level was an important transparency measure, just as the 2020 AI Principles are as well. With these documents and others, we have, really, the building blocks to work with allies and beyond because we have to get to the "and beyond" part as well on corporative measures such as establishing robust T&E and V&V standards for such practical measures, such as tripwires.

Paul Scharre: Thanks. Well, that's very helpful, again, to hear the international perspective as well and certainly let's not feel constrained to only the issues of Chapter Four. If there's other things you want to leap on that are not related to the report, feel free. One of the things that I think is interesting about the report, is there are some specific ideas that they tee up in terms of confidence-building measures, as you mentioned Kerstin, that it's able to get beyond some of the hand-waving of, like, we should think about ways to mitigate these risks and talk about particular ideas.

Paul Scharre: I want to tee up for the group the first recommendation that they have when it comes to confidence-building measures, which they talk about having the U.S. clearly and publicly affirming existing U.S. policy that only humans can authorize the deployment of nuclear weapons and then look for similar public commitments from Russia and China. I'm interested in your thoughts and reactions to that. Is it a good idea? Are we lucky to get anywhere with it? What's your take? Jump off for anyone who's interested.

Helen Toner: I mean, I'm definitely a fan of that recommendation. I think it's an absolute no brainer right now, that there's no way we could ... Maybe 10, 20, 30 years in the future this is something that it would make sense to revisit, maybe, maybe. But right now, it's a complete no brainer that you don't want to have any of these kinds of systems anywhere near your nuclear C3. I think that's something that the U.S. could easily affirm. The former Director of the Joint AI Center at the Pentagon, General Shanahan, has said this, but you know he's one person at his rank; to have it affirmed more loudly and by more senior personnel, I think would be—is clearly a good move.

Kerstin Vignard: I completely agree with Helen on that. I mean Paul, you really hit it on the head, this is confirming existing policy, right? That should be said loud and said proud, and up the stakes for others to make similar commitments publicly. It's a super CBM to make public unilateral declarations and then use that as a way to pressure our friends and allies to help develop that normative framework around it as well.

Michael Horowitz: Yeah, I think it's not just a good idea, I'll make it uninformatively positive here, but it's good strategy. I mean, we never like saying that we won't do things. That's just like not something the U.S. really does all that often. It's not just General Shanahan. General Rand a couple of years—there's several American generals, like prominent American generals—have expressed serious concern with the notion of sort of AI integration in critical nuclear systems. Part of that is about technology risk. I think part of that's about trust in America's existing second-strike capabilities. In that case, you know, why would you take the risk?

Michael Horowitz: In that case, there's a real opportunity for the U.S. to be a leader in this case, which, to Kerstin's point, I think then maybe puts a little bit of pressure—even if it's only a little bit—puts a little bit of pressure on other countries to maybe make similar statements and helps sort of set the standard. From a strategy perspective then maybe makes it easier down the road when other countries want stuff the U.S. doesn't want to do.

Helen Toner: Yeah, what this reminded me of—to your point Mike about this saying, we won't do things is not something that happens here very often—reminded me of President Nixon saying that doing the unilateral commitment to get rid of the U.S. offensive biological weapons program and that obviously looks fantastic in retrospect. Obviously, a good idea.

Michael Horowitz: Right, who needed bioweapons? You had nuclear weapons, there was no point to them anymore, it's pretty easy.

Helen Toner: It's automated nuclear weapons, not human nuclear weapons.

Paul Scharre: Yeah, I mean I think one of the things that I've thought about with this—and maybe this is unfair—but it seems to me like just a bare minimum threshold for when we think about things that you're concerned about with AI risk, it seems like an easy one. In particular, it seems to me that if we're not able to get other competitor nations to publicly, in a political statement that's not legally binding—that's not a treaty—to sort of say like, “yeah, we agree that we shouldn't allow machines to make nuclear launch decisions,” then I don't know how we're going to reach an agreement on more difficult challenges. I would hope that's a bare minimum that we can get countries to agree upon. It seems like at least a good place to start. Other thoughts on things that jumped out at each of you in terms of the recommendations, or some of the risks that the report highlighted that you want to highlight?

Michael Horowitz: I would just call out one thing, circling back to something that I think Helen had mentioned and thinking about the contrast between some of the chapters. I think that there is something really interesting here about what competition means and whether we think about competition as positive or negative in a way. You know from an international security perspective or international stability perspective, we almost inherently think of competition as a negative thing. Because competition is what generates arms races, it creates risks, it makes conflict more likely, it can make accidents more likely, et cetera.

Michael Horowitz: But competition when it comes to sort of broader technology development could be about competition for economic reasons, for things that are not about the use of force. So, I think part of my read on this report is it's actually pretty nuanced in the way that it thinks about competition, because in some ways it's pointing out the—I don't think they actually explicitly say it this way—but not just the need to compete, but the benefits of competing when it comes to AI leadership more broadly, and then trying to buy down what some of those risks could then look like. What are inadvertent consequences of that applied in the military arena specifically, as you're trying to capture broader benefits from AI development—from the perspective of the report since it's an American-focused report, from the perspective of American society, America's influence in the world, U.S. technology leadership, et cetera.

Paul Scharre: Helen, we've chatted before about sort of the confidence-building measures, CBMs as we've been calling...people call them. For those who've maybe not, we've been using the acronym CBM—confidence-buildings measures—right? We've been chatting before about that in previous discussions and workshops and you've had a lot of contributions to this. I'm interested in other than the nuclear one, what you see as maybe some of the most fruitful potential confidence-building measures for reducing or mitigating AI risks?

Helen Toner: Yeah absolutely. There are a couple that are in the NSCAI report that I liked a lot, and I was glad that they were included. One of those was direct military-to-military dialogues. So, the report I think recommended that these kinds of AI stability issues be included in the existing strategic security dialogue with Russia, and that a similar type of dialogue be set up with China. I think that's a great idea. I think trying to talk about some of these crisis dynamics and how to manage and mitigate them and also talking about intentions in those kinds of fora is really valuable.

Helen Toner: Another one in the report is working towards testing and evaluation standards for these settings. Right now, it's really not clear how you would go about testing and evaluating a machine learning-based system because they're just so different from what we have—the types of autonomy that we have in play at the moment. Then, one other that I'll mention that I think was in your report, Paul and Mike, is sort of another unilateral one, which is just about signaling and stating clearly about the difficulties of these T&E and V&V problems.

Helen Toner: I'm actually interested if any of you have a take on why this doesn't get said more loudly. It sort of feels to me like there's this open secret, which is all these machine learning systems that are doing great, that are fantastic in recommending your next Netflix show, they're really good at labeling your photos, but they're absolutely terrible if you're using them in any kind of sensitive setting, where the outcome really, really matters. Pretty much everything battlefield-related in such a setting, why is it so difficult for us to say, we're a long way from being able to do this?

Helen Toner: Is it just that DoD and other militaries want to signal competence and signal that they're moving ahead? Or is it that this just hasn't filtered through and isn't something that senior leaders feel confident enough to say yet? I'm interested in anyone's take. Looks like Mike has thoughts.

Michael Horowitz: I mean, I suspect all of you actually might disagree with me about this. But I think that that is better understood in DoD than it's commonly appreciated. I think you see that revealed through the way the U.S. hasn't deployed a lot of wild stuff that people would sort of think is unsafe. Like, if the best evidence we have for how responsible the U.S. military is being is what it's actually deploying, then I think the U.S.—that's almost an explicit acknowledgement about the limits and ability to do testing and validation and verification of some of these systems.

Michael Horowitz: I mean, I think that if the question is why it isn't louder, I think that gets to more—I think that's almost a question for the media about what people wish to report and talk about as much as anything. But I think, within the community of people that are thinking about what investments should look like, I think some of these issues are, I would say sort of reasonably—I mean, well understood is probably putting too fine a point on it, of course there's going to be variation, but I think it's not like people are unaware of the challenge.

Michael Horowitz: I think the question becomes more like, do you believe in the way that—essentially, do you believe in the system—such that if you can't design effective testing and validation then something just won't get deployed. So, I think in some ways it's if you believe in the system then you think that the Defense Department's likely to behave responsibly about this. If you don't inherently believe in this system, you think that they're likely to sort of take short cuts because you think there are these hard problems that haven't been solved yet, which is true, just to be clear. Like not put you back in the substance at all.

Helen Toner: Yeah, that make sense. I mean, I guess the thing that I find a little surprising is that this isn't called out more in things like... To go back to the NSCAI report, they say one after another—they first say, "the U.S. has procedures in place that will take care of this, don't worry about it, we're not going to deploy unsafe things." Then they say, "other countries don't, and we're concerned about that." Why not say, I don't know, I feel like there could be more emphasis there on, "and therefore, countries that don't have these procedures in place risk putting systems on the battlefield that are going to backfire and blowup in their face and cause massive issues." I guess they sort of say it, but they, I don't know, it doesn't feel like to me like it comes out and says that as plainly as it could be.

Michael Horowitz: I mean yeah, I guess I feel like the report says that, but it probably, it could be foregrounded more I guess.

Kerstin Vignard: Yeah, I think that the issue of the hype around it, I think Mike you're probably right, there's a lot of nuance to understanding inside. It doesn't carry very far though, a lot of times. It gets kind of the nuance, every successive repeating of it, the nuance gets more and more lost. I mean, there's a lot of nuance in the Commission's report, it's 700 pages long right? Minus the annexes. So, what we people hear is the hype, and that is something—having, I think, a posture that says, we believe for these reasons that this is an appropriate use of this technology given its current state, right?

Kerstin Vignard: That is not saying we're the best, we're the best, we're the best. It's like, this is just a reality check and that goes to that tension we often see in, for example, the LAWS discussion in CCW, of this technology is completely revolutionary and it changes everything and on and on and on. It's completely familiar, we don't need anything new, we don't need to talk about new frameworks, and there's the tension between those two statements. Usually, they're said by the same delegation, one right after the next: "This is amazeballs, but we don't need to do anything new." There's not a lot of calling out on that, and there's a lot of cognitive dissonance that happens when you hear those two statements at the same time. So, I think more nuance is important on these issues and that's across the board, whether it's AI-enabled weapons systems or AI in other parts of our lives.

Paul Scharre: For those following along with bingo cards who had "amazeballs" on it, I think you just won. Thank you, Kerstin. No, I think this is great. I think you raised a good point Helen, which is I consistently see here this disconnect between people who work on technical, machine learning issues and those in the defense community—how they talk about some of these safety…particularly the safety concerns. I hear more concerns about security risks and hacking from the DoD side than I even knew about the safety issues. But, I mean, that was one of the responses that I saw from folks about the report, was like, "how are you going to use machine learning systems in these ways? It's not reliable enough." I think at the technical level, like the engineers, probably understand those challenges at DoD. But I think the tone is clearly different and the way that senior leaders talk about it is clearly different, and I think the messaging, my take is, at least the messaging for senior leaders in the Department—and the Commission report kind of reflects this—is largely like jam your foot on the accelerator. It's like how do we do more? How do we go faster? It's like yeah, yeah, yeah, T&E, we'll figure that out, let's move forward. I don't know, I'd be interested in others' reactions.

Helen Toner: Yeah, I think that's similar to my take and I feel sort of ... I guess the thing I feel confused about is why there isn't a clear ... It seems like it would be in the U.S. interest to say more clearly, “hey Russia, hey China we all see these systems aren't ready yet, right? Like we all see that these technologies are not going to be able to be deployed in a useful, reliable way any time soon.” I think the messaging that is being done is a little different from that.

Paul Scharre: So, we got a question that's come in through the chat here. What would a verification and compliance regime for confirming no AI or machine learning use in a nuclear decision-making look like? How would you verify AI's role in strategic nuclear command and control? It's very different than verifying hardware components or other types of WMD. Interested in folks' take on that, how would you implement that in practice? Mike, what do you think?

Michael Horowitz: I think I'd say two things about this. One is that there's an extent to which there are probably international benefits to be gained, even if there's not verification. To go back to an example that Helen gave before—from the Biological Weapons Convention—there's not a verification regime associated with the Biological Weapon's Convention, I think to the chagrin of many, including me.

Michael Horowitz: Certainly, I think that has maybe made violations more likely at some point or another. But creating that strong norm against the use of biological weapons we've come to regard as important. So, one could imagine some sort of international declaration surrounding nuclear weapons and AI, even if it didn't include verification, is actually still having, I think, some important value.

Michael Horowitz: I mean, I think verification there would be extremely tricky and frankly, there's not a ton out there that explores some of those things. Let me toss it to Helen because one of the only things I've read about this—this report—that came out from CSET two weeks ago, that explores some possibilities for how you could essentially have inspectors plugging into systems to verify sort of what's there or not there. Now the political will to have people do that is a different question.

Helen Toner: Yeah, I mean, I totally agree that I think we need not assume that we need to have these kinds of verification mechanisms in order for these kinds of declarations to be useful. At the same time, I think—Chapter Four of the NSCAI report, to come back to it yet again—actually had as a specific recommendation to pursue technical means to verify compliance with future arms control agreements pertaining to AI-enabled and autonomous weapons systems.

Helen Toner: I think that this, as a technical project—so as you said, CSET put out a report about this a couple of weeks ago, which is very much an initial look. I think right now, the answer is we don't have great ways to do this. But I think as a technical problem and a really interesting technical challenge for the next—for the coming years—my understanding is that there seemed to be potential things that one could explore here. There's also a super interesting paper called "Structured Transparency" that came out recently on a whole category of approaches. The idea there is that, not to go off on too much of a tangent, but when you're using computers to check things, there are ways to use the computers to check the things that don't actually reveal the things to the humans looking at the computers. So, you can think of this sort of as analogous to the way a sniffer dog in an airport sort of lets the border agents inspect someone's luggage without actually looking inside of the luggage.

Helen Toner: So anyone who's interested in this could look at this in the CSET paper on verification or also look at this other paper on structured transparency, which is looking at some of these technical approaches that let you verify properties of a given system without necessarily looking at it. Those approaches also have relevance for privacy and for other data sharing issues as well. It's sort of a bigger space that I think is really super interesting for the years to come.

Paul Scharre: Kerstin, any thoughts on this issue of verification and compliance?

Kerstin Vignard: I completely follow on what Helen and Mike have just said. I think Helen also kind of left it on the right note, "for years to come." So we can't start with—we can't make agreements—unless they are verifiable, or we can't take steps towards risk reduction if it's not verifiable, because we will be waiting for years, and that's not a good place for us to be. So, I think it's important to work on verification issues, and there's some really interesting work going in that direction, but that can't be what holds us back. Again, that brings us back, perhaps, to confidence-building measures and their importance, and what can be done that's practical now, while we are working in the research and scientific domains to get us towards more verification-type of arrangements.

Paul Scharre: So we've been talking a lot about some of the recommendations in the report. I guess I'd be interested in folks' takes on things that you think are significant that the U.S. could be doing, either unilaterally or with others, to mitigate AI risks that maybe weren't included. Maybe things that were left on the cutting floor or confidence-building measures that might be worth exploring?

Michael Horowitz: Let me start by saying... In some ways, while I think Helen and I might disagree about the extent to which the report emphasizes, like some of the risks in current AI applications, I think she's certainly right. Like the disagreement is what the report signals, not whether that's a good idea. So to the extent that, if essentially, if Helen's read is right, then that's something that was left on the cutting-room floor, that the U.S. should thus be promoting more. Because that would be a good idea because it's true.

Michael Horowitz: I don't think the U.S. would be disadvantaged by making clear the limits of both existing—of both testing and verification—and I mean, look, not to be like too professory about it, but like to the extent that one is a Clausewitzian and believes in the fog of war and—sort of friction is inherent in warfare—then the notion of an algorithm that actually will know everything, could have taken and been trained on everything that could possibly happen in a multidimensional battlefield, like...kind of challenging.

Paul Scharre: So, another bingo win for those who did not have Clausewitz, but Clausewitzian...very important. So yeah, I mean, I don't know, Helen a response?

Helen Toner: Yeah, I mean, agreed. I think another element that I'm pretty sure was one of the chapters of the report—but I have not looked at all the recommendations and all of the chapters yet—that the U.S. could do, it's pretty straight forward, is just investing in the research to try and overcome these problems, and to the extent possible, collaborating with allies and partners, even to the extent possible, collaborating with competitors and adversaries. The go-to example here is ... In the Cold War, the U.S. was handing over technology that secured nuclear weapons and made it possible to... anyone know the name of the thing I'm talking about? It's totally slipping my mind.

Paul Scharre: Permissive Action Link is that what you're...?

Helen Toner: Exactly.

Paul Scharre: Yeah.

Helen Toner: Thank you, yeah. The U.S. developed this technology that let you secure who was going to be able to make use of a nuclear weapon and then they just handed that to the Soviets because it was obviously in everyone's best interest for rogue Soviet operators not to be able to use their nukes. So, I think there are potentially analogous advances that could be made in how we build AI and machine learning systems that are more robust and reliable, and certainly making sure that our allies and partners have access to as much of that work as we can, and drawing on their best scientists and researchers as well. And then, depending on the specifics I guess, also trying to make that available to competitors and adversaries.

III. Audience Q&A

Paul Scharre: We've got a question coming in from the group about explainability, a topic that comes up quite a bit in machine learning and particularly has relevance in really any sensitive application, but certainly in national security ones. Interested in folks' take on how significant explainability is when we think about some of these applications?

Helen Toner: I mean I can jump in to say very significant. The thing to understand here is—essentially the way that these sort of cutting-edge machine learning systems work, these deep learning systems in particular, or deep neural networks, it's not that we don't know how they work, kind of mathematically, it's that if you look at how they work mathematically, it's sort of millions or even billions of numbers that get multiplied together in specific ways. So we can look at those individual numbers and that doesn't actually tell us what is going on for this system. You can think of this both in terms of autonomous systems—so an automatous system goes and does something, and you want to look afterwards, why did it do that? Or you can think of it in terms as well of decision support systems, so you know, maybe a pilot has a heads-up display and it makes some recommendation for what to do next, or it tells them that it's identified something in the surroundings. Can the pilot understand why does the system think that? Should the pilot believe it, or should they trust their own judgment?

Helen Toner: Again, the short answer here is basically, we currently have no good ways to set up these cutting-edge deep learning systems, such that a human can look at them and really get what was going on. This is an ongoing area of research, there's some super cool work that's come out, even in the past month or two. But it's a long way from being deployable and useful and reliable.

Paul Scharre: I want to follow up on the issue that we talked about earlier about technology risk, the maturity of the technology. We talked about it in the context of signaling and how the U.S. should message this. But I'm interested in everyone's take on—internal to even just how the United States approaches AI and machine learning technology—do you feel comfortable with the processes in place to ensure that the systems are mature and what's coming out is reliable and trustworthy, or do you think there's work to be done there?

Michael Horowitz: I think there's certainly work to be done there, but I would say, as an earlier answer suggested, on balance, I guess the way I would say this is, I certainly believe in how the U.S. is doing it, more than my understanding of how anyone else is doing it. That actually relates back to the explainability point that Helen just made, which I agree with entirely. I would just add, I actually think explainability is an important thing here since from a technology adoption perspective, if senior leaders and the operators that have to employ these systems don't understand how they work and can't be persuaded that they can understand how they work because the systems aren't explainable, that makes them less likely to be used anyways.

Michael Horowitz: So, I actually think solving—well not maybe solving—mitigating, addressing, whatever, the sort of the explainability issue is part of what the pathway forward is to safe and reliable uses of AI for the Department of Defense. Because that's what will make it easier to explain to broader sets of both the senior leader and operator communities of what these capabilities are, and what the risks are, in ways that are accessible to them, which makes them more likely to behave responsibly. Essentially, if you want to decrease, say, if you're worried about automation bias and people thinking about sort of an algorithm as like magic pixie dust, then they don't realize it's not like a calculator. The explainability and thus the ability to more clearly articulate risks based on that, and what failure modes might look like, could thus make safe and reliable adoption easier as well.

Paul Scharre: Kerstin, Helen?

Kerstin Vignard: I was just kind of just thinking if there's any lesson that we could take from the cyber negotiations that have been underway for 15 years in the UN on digital technologies, you mentioned signaling. I asked myself, rather than kind of looking for there to be a new international arrangement and new agreements and things, let's really look at the norms and principles, the CBMs, the legal regimes, that we've already agreed on—or the states have already agreed on—on international peace and security and digital technologies, and see where we can use those to get a leg up.

Kerstin Vignard: Particularly, I think that includes Article 36 weapons reviews. Now that's weapons, means and methods, and really pushing forward on the means and methods part, I think it helps us with signaling, and we can loop that back to T&E and V&V. The closer work today with the technical community, more investment with the research community, is really going to help policymakers understand what would be the bar for T&E and V&V, for AI-enabled means and methods—not just for weapons, but the means and methods—and that would probably be an important signaling function in the international community as well.

Paul Scharre: Do you want to maybe just briefly explain to people what Article 36 reviews are, what that means? Some folks I assume will know, but maybe not everybody on call.

Kerstin Vignard: Article 36 of Additional Protocol to the Geneva Conventions, where states who have signed up to the protocol—please correct me if I'm wrong at any point, Paul—go through weapons reviews to test and ensure the legality of their weapons. We talk about them as weapons reviews, but if you go and look at the actual language, it talks about weapons, means or methods, of warfare.

Kerstin Vignard: There's particular challenges when we talk about digital technologies in Article 36 reviews, which are not transparent processes obviously. There are particular challenges with digital technologies, and particularly when we talk about learning systems. Testing and going through an Article 36 review with a system that may evolve on the battlefield brings up particular challenges. These are, I think—people have been thinking about this—but I don't think there's a lot reassurance coming yet, partially because we don't have the technical side of the house in order yet, and it's still a lot of open research questions. But as Helen mentioned, a lot of interesting progress happening there too.

Paul Scharre: So we've got another great question coming in from the group that I'm just going to read aloud to folks. So—and this is a common question that I hear—so I really like... thanks to someone for submitting this. Have the speakers seen much evidence of similar policy debates in Beijing or Moscow around these same topics? The person notes neither Xi nor Putin have an equivalent commission offering outside and public advice, but are there policymakers or experts in China and Russia who are calling for their governments to explore confidence-building measures and to mitigate the risk of military AI? What's your take on what might be going on, perhaps even behind the scenes in other capitals?

Helen Toner: I've seen a little of this in China. Paul, I'm actually interested in your take, I'm pretty sure you would have interesting stuff to add here as well. I think it's always difficult for Chinese representatives to—they're often more concerned about speaking out of turn at international events than their perhaps American counterparts. The event where I've seen most discussion of this was an event at the National University of Defense Technology in Changsha in 2018 which was a mostly Chinese language event and most of the participants were Chinese. There were many different presentations about AI and global security issues and maybe two or three of them touched on these kinds of stability issues and made kinds of recommendations along these lines. So, that was a place where I felt like I got to get a little bit of a peak into more of a domestic discussion and these issues were on the table and I was glad to see that. Though they weren't as central and certainly not as public and splashy as the NSCAI report.

Paul Scharre: Yeah, I mean I would certainly agree. I think there's obviously a big difference in terms of public messaging and there's just different audiences, right? I mean Russia and China don't worry about some of the domestic messaging and audiences that folks do here in the United States. I would say that I've been a part of track II dialogues with Chinese military experts over the last... I guess year and half, going on almost two years now. I found much more similarities than differences in how Chinese military experts studying AI talk about these issues, very similar to how folks in the U.S. do.

Paul Scharre: There's a general skepticism of some kind of legally binding measure that might tie your hands in a competition, but I think a strong awareness of the risks of the technology, the limitations of the technology. I think one important asymmetry is the legal and ethical issues I found just have less salience when talking to folks coming from China. Whereas when talking to folks in the United States or in Europe, where the things that people might come with first is, "well what about the law, what about these ethical concerns?" That's not what foregrounded with Chinese experts, but they are deeply concerned about retaining human control over the technology.

Paul Scharre: So I'll often hear, sometimes people in the U.S. say things like, well you know our adversaries don't share our morals and our values and so they would be unconstrained. I think the first part of that thing might be true—that they might have different moral and ethical perspectives, clearly on a number of issues, particularly with the Chinese Communist Party—but when it comes to constraints, there are other practical constraints about militaries wanting to have weapons that are controllable, that also exist in other systems as well. Certainly, in authoritarian systems, there is a real strong desire for a strong and tightened degree of control by the political leadership over the military forces and what they're doing.

Paul Scharre: But I think these kinds of ... There's also... and this is just worth mentioning here—I think most experts who track these issues are well aware of it—just a huge asymmetry of knowledge between the U.S. and China, where there's a lot more understanding, I think, that people have in China about what is going on here, than vice versa, and more work to understand how other experts who think about it is valuable. We are coming up on the end of time and I did not mean to filibuster there at the end. Let me turn to each of you to give you 30 seconds to jump in with any parting thoughts. Kerstin, you're up first.

IV. Closing Remarks

Kerstin Vignard: I guess my parting thought is—while I am a multilateralist at heart, and I believe in multilateral... the importance of multilateral discussions—we need much more engagement from a multi-stakeholder community, not just states, and particularly technical communities, helping us to understand how to manage risks, in order that we can get the benefits of AI-enabled systems.

Paul Scharre: Thanks. Mike?

Michael Horowitz: Sure, I think there are a lot of opportunities for the U.S. to be a leader, both in AI, which I think would be good from an American economic and national security perspective, and in AI safety, and in decreasing the risks to international security from more dangerous uses of AI, be it a miscalculation, inadvertent escalation, et cetera. I mean, there's a lot of precedent for this. I mean, as you know, I think the better analogy for AI is something like a general-purpose technology, and so the right era to be thinking about in general is more like late 19th, early 20th century.

Michael Horowitz: But if you think about the Cold War broadly, the U.S. simultaneously promoted varieties of arms control and confidence-building measures while pursuing technical, economic, and military leadership. It doesn't need to be one or the other. That, I think, is in the report—maybe could have been clearer in the report—but it's something that I think is going to be really important moving forward. That we can, in fact, walk and chew gum on this at the same time.

Paul Scharre: Thanks. And Helen, you get the last word.

Helen Toner: Sure. I mean I think Kerstin and Mike said it very well. Maybe to try and combine Kerstin's note about the need for engagement from technical experts and Mike's on the U.S. leading on some of these safety and standards and testing issues, one thing we didn't get time to talk about today is the NIST process that's getting started on building a framework and trying to build some standards for robustness, and reliability, and interpretability, and privacy. So, I'll just say that I'm excited about that process, I think we're going to need some really great minds to get involved with it—to get it to go as far as it can. But I think that's another thing to watch.

Paul Scharre: Well, thank you. Thank you all for joining us, thank you for your insights and thanks to everyone who was here for this discussion and thank you all. Take care, have a good day.