Challenges and Opportunities for AI in Military Systems

[00:00:00] Welcome and Introductions

Sheena Chestnut Greitens: Welcome to *Horns of A Dilemma*, the podcast of the *Texas National Security Review*. I'm Sheena Chestnut Greitens, editor-in-chief of *TNSR*, and I'm here with Dr. Ryan Vest, our executive editor. I'm pleased to have joining us today, Michael Horowitz, author of the article, "Artificial Intelligence and the Future of Strategic Stability," which appears in the recently published special issue of the *Texas National Security Review.*

Mike Horowitz is Director of Perry World House and Richard Perry Fellow at the University of Pennsylvania. He's also a senior fellow for technology and innovation at the Council on Foreign Relations. From 2022 to 2024, he served as deputy assistant secretary of defense for Force Development and Emerging Capabilities, and director of the Emerging Capabilities Policy Office at the Pentagon. Mike, welcome to *Horns of a Dilemma*. It's great to have you on the show.

Michael Horowitz: Thanks so much for having me. Super excited.

[00:00:51] AI and Strategic Stability

Ryan Vest: Mike, I love this article, and in it you frame the article around what you call the intersection of time, uncertainty, and confidence in capabilities, arguing that the two competing perspectives about AI and strategic stability may both be correct at different points in the technology's life cycle. What are these two perspectives and why does this framework sit at the core of your argument?

Michael Horowitz: Sure. I mean, I think that *a*, when we talk about the way that emerging technologies will influence international politics, we're often prone to making sort of universalist kinds of arguments, like, oh, like, AI will be a disaster for the international security environment or, you know, quantum, oh, that means, like, no more cybersecurity ever, when the reality is that it can depend on, like, all the same factors that influence military power in other

contexts, which is shaped by essentially the relative balance of power—like, what does one side have versus the other side have?—and frankly, like, how mature is the technology? And in thinking about artificial intelligence and strategic stability, you've seen a couple of different perspectives laid out.

The first, which is, I think by far the, like, minority sort of viewpoint, is that you could imagine a world where AI enhances strategic stability. If you think about, like, "What are algorithms good at doing?"—a nd that is, you know, processing information faster than people—then things that algorithms might be really good at could include something like early warning surrounding a potential nuclear attack.

And given that we imagine, for example, that decision-makers make better choices when they have more time, since that AI buys decision-makers more time, maybe there's an argument for why it would enhance strategic stability.

More people, I think, have made the opposite argument, which is that advances in artificial intelligence would actually be really bad for strategic stability. They point to the sort of error proneness of algorithms. Like, anyone that's ever had, you know, ChatGPT throw an incorrect answer, like, back at them or, like, tell them something that, like, definitively wasn't true— that there you might have overconfidence in algorithmic accuracy or what's called automation bias. You could frankly have the opposite, algorithm aversion, where people don't trust algorithms as much as they should.

You have a range, essentially, of different kinds of biases that could lead to the integration of artificial intelligence in dangerous ways that then cause, you know, lots of problems for the international security environment. Where time I think factors in here is that, you know, over time of, like, "What do we do?" we tend to figure out how technology works and figure out how to use it effectively and in ways that are safer and more responsibly. Whereas, like, really early on when people don't know what they're doing and where expertise is much lower, like, things like responsible use with emerging technologies becomes a lot more difficult.

So it's easy for me to imagine, you know, for example, that the introduction of AI into something like nuclear command and control, not saying it's a good idea, could be destabilizing maybe early on in a period, but then as everybody kind of understands how to use it and is better calibrated in what algorithms can do and not do, then maybe it becomes more stabilizing. The point being it's not just, like, a linear story.

Sheena Chestnut Greitens: So that's actually one of the things I found really fascinating in reading this article, is this idea that psychology and organizational behavior are really going to shape how AI conditions nuclear risk and that those aren't going to be linear or flat over time, right?—the effect of those factors.

[00:04:12] Misconceptions and AI History

Sheena Chestnut Greitens: So I want to come back to that in a minute, but maybe before I do, you also walk readers through the history of artificial intelligence in this article in a really accessible way, and I found that really helpful. So I wondered if, for the students and for listeners who are trying to get a handle on what AI is and all the ways it might affect debates in national security, how you see persistent misconceptions about AI making it difficult to have a conversation about the nuclear implications. And, maybe, are there lessons from earlier periods of AI's development that we should be looking at instead to frame the discussion about it today?

Michael Horowitz: Yeah, absolutely. I mean, I think one of the things that's fascinating is artificial intelligence is in some ways, like, the thing that we can never actually achieve. And we see this over and over again, over like a 60 to 70, you know, some odd year period. And that if you think about AI at the core it's computers being able to do things that used to require human intelligence.

You know, at first people said, "Oh, you know, if a computer could win at checkers, you know, that'd be like a really smart computer. That would be artificial intelligence." Then it turns out that, like, that's actually not that hard to do once you have computers and you can program a little bit. They were like, "All right, all right. That was too easy. That's not artificial intelligence. If a computer could win at chess, that would be, like, a really smart computer. That would be artificial intelligence."

And then it turns out, that's, like, really hard math, but that's a solvable problem. And then they're like, "All right, well, that's not really intelligence. That's just, like, solving an equation." It's like, oh, if a, you know, algorithm could win at Go, like, that would be artificial intelligence. And then of course, you know, you have now, like a decade ago, essentially this, like, alpha Go moment where an algorithm beats one of the world's leading Go players.

And

Ryan Vest: then

Michael Horowitz: people say, "All right, well, but, you know, Go's like mostly equations. It's not like that that hard." And so, you know, in some ways you can imagine this, if you're a sports fan, like, the goalposts are always moving on what constitutes artificial intelligence. And what's happened especially is, as the frontier of artificial intelligence advances and people start talking about things like artificial general intelligence or artificial super intelligence or, like, all of those, like, kinds of things—which the article doesn't actually, like, get into, like, that much—one of the things that we forget is that there are all these things that we, like, don't even call AI anymore that involve, like, even old school algorithms or machine learning—like, those kinds of things— that actually have been now embedded into military systems for decades and that are already sort of shaping how militaries behave and think about how to use this technology from a national security perspective, which, like, on the one hand frankly, should be reassuring.

This is, like, not the first rodeo for, you know, like, big national militaries like the United States in, like, thinking about algorithms. Like, on the other hand, to the extent that, you know, you think about artificial intelligence as a general purpose technology the way that I do, and it's impacting everything. And we're in a world where, like a decade from now, you could have AI embedded in every military system, frankly, like, nuclear and non-nuclear. The question is, like, where and how and for what use cases. And, like, that I think is the question that is often missing from the conversation. What kind of algorithm makes sense probably depends sometimes on what are you using it for and how high leverage is that use case?

Like, the regular, you know, like a ChatGPT or something or like a Claude trained on, like, all of the data from the internet is not going to help you identify whether, you know, the Russians have launched a missile faster. That's more of like a bespoke data question, like, for example. And so, like, what data, what algorithm, what use case, like, that all really matters a lot in trying to unpack this question.

[00:07:50] Simulated Data and Teaming

Ryan Vest: So I want to jump in on that a little bit, because in the article, a recurring concern that you come up with is that many AI systems, especially those in the nuclear domain, need to be trained on simulated data, given that there are no real world examples.

Michael Horowitz: Which is good. Just to be clear.

Ryan Vest: Right.

Michael Horowitz: I'm anti-nuclear war.

Ryan Vest: Yeah, we don't want that. But with stakes that high, when you talk about nuclear war, how should policymakers think about deploying systems that we can never thoroughly or fully train through experience?

Michael Horowitz: I mean, that's true. Also, it's inevitable, and it applies to humans today. If you think about, like, if the question is, "Could we use an algorithm to detect a missile launch?" and we need to use simulated data to train the algorithm. If the alternative is, like, a human who's sitting there like bleeding out of their eyeballs because they're, like, standing there, like, looking to see if there's a missile launch, like, what data are we using to train them? It's that same simulated data.

So I think that there is a real question about the extent to which simulated data can mimic reality. There are a whole bunch of debates out there in the AI world on that context. If what we are talking about in, like, an early warning—I've, like, made it as easy as possible for me now, and so, in answering this question—but if what we are talking about is a missile launch, that's actually like a relatively discreet, like, kind of thing to be able to, like, train an algorithm to look at. And so I feel, like, relatively more confident in that, but there's still risk.

But this also gets to a question that I don't think we ask enough when it comes to the integration of AI into military systems, which is, "What was doing the job before, and how good was it?" And that, like, to me, this is all relative. Like, if you tell me that there's, like, an AI early warning system that's going to be like super error prone, then obviously I'm not going to be excited about it, but if you tell me it's super error prone, but, like, actually the humans that were doing it before were more error prone, that now raises like broader questions about like systems and processes that maybe we should also be solving.

This is also why I think the notion of human-machine teaming, when it comes to AI in the nuclear context, seems like really important, and that the integration of AI into intelligence, surveillance, and reconnaissance for conventional operations by militaries, like, that's inevitable. Like, that's happening. There are timing and implementation questions, but that's a thing. And to the extent that nuclear command and control relies in part on some of the same sensors that, say, a military like the United States or China or somebody else, you know, would use in the context of a conventional conflict, that means, even if you're

trying to exclude AI from, you know, some bespoke nuclear systems, you still are going to have a degree of integration in existing systems, which [is] why, to me, what we're really talking about is sort of human-machine teaming, human-machine integration.

[If] what we want is accuracy, then what we want is the best of what machines can do with the best of what people can do. And that requires, like, new kinds of then training and developing standard operating procedures. But, like, that's a thing militaries can do.

Sheena Chestnut Greitens: I think this is a really important point, right? We have to be really careful not to assume a fallible AI and an infallible human because we just know that's not the way the world works.

Michael Horowitz: I don't know if you know this Sheena, but people are not good at stuff sometimes.

Sheena Chestnut Greitens: Really? I didn't know. No, so this question about human-machine interface and human-machine teaming makes me want to ask about a couple of questions or biases that your article talks about that could come up at that interface, right?

[00:11:17] Automation Bias Explained

Sheena Chestnut Greitens: And so one of the things you walk us through in the article is this idea that AI -enabled early warning systems could either reduce or increase nuclear risks and that one of the things that could shape the level of risk is automation bias.

So can you just tell our listeners, you know, first, what is automation bias? How does it change the role of human judgment in crisis decision-making? And then how does this shape the broader context of nuclear command and control?

Michael Horowitz: Yeah, absolutely. So automation bias is when people trust algorithms more than they should based on the, like, known accuracy of the algorithm. So imagine an algorithm that is 80% accurate at whatever the task could be, whether it's, like, determining whether something is a cat or determining whether something is a missile launch or, like, whatever it is. If it's 80% accurate, that means it's wrong 20% of the time.

But if you take that 80% accuracy and people treat it as 85%, 90%, 95% accurate, that's what automation bias is—when you trust algorithms more than you should, based on the accuracy of those algorithms. What it leads to is what's called cognitive offloading, when people essentially take the keys out of the car and, like, hand over control over to the algorithm.

And we've seen accidents happen for this reason before, like, out there in the regular world. You know, we see in varieties of airplane crashes where autopilot has been at issue. Like, one of the reasons why you sometimes have these issues is that pilots are like, "Well, but the autopilot says that we're fine. Like, how could there be a problem?" when, really, there are all sorts of other indicators that suggest that something is wrong. But they have learned to trust the autopilot and trust the autopilot too much based on how accurate it is.

We're not, like, condemned to fail because of the risk of automation bias. What it suggests in some ways is that education and training and, you know, procedure is necessary to hedge against it in some ways. And that becomes more necessary as algorithms improve more, because if you have algorithms that are like a coin flip, then, I mean, that kind of raises the question about, like, why you're using them in the first place.

But, you know, imagine you're, you know, you're using your coin flip algorithm, it's going to be pretty clear that it's a coin flip algorithm. And so the degree to which you rely on that when making decisions is going to be pretty limited. But if you have algorithms that maybe nine out of 10 times are right, then the instinct to trust that algorithm can obviously grow, which then makes it a lot more difficult in some ways to, like, challenge that judgment, when it comes to decision-making.

What can really influence this? And I published a paper with a great co-author named Lauren Kahn about two years ago in, sorry, a different journal, about automation bias that showed, essentially, a version of the Dunning-Kruger effect in play, where at basically no knowledge about AI, people were kind of algorithm averse. They're like, "I don't want anything to do with that algorithm. Like, that seems, like, kind of, like, wacky and dangerous." At low levels of knowledge, that's when automation bias became most likely. People became overconfident, essentially— like a little bit of knowledge can be dangerous, like that kind of story.

And then at higher levels of knowledge, then people started to get better calibrated. And, you know, better calibration is when you reduce the risk of automation bias, and that's frankly what we're going to want to the extent that

you have AI integration in the nuclear enterprise at the sort of early warning kind of stage, which I just said in the answer to the previous question, kind of is inevitable because of conventional sensors. Then we need to get to that point of calibration, which means now there's, like, work to be done on AI literacy, on AI training, on, you know, like, doctrine, et cetera.

Ryan Vest: So this is kind of interesting. The theme I'm starting to pull out of this is that the interface between humans and AI is maybe the most important thing that we need to start focusing on in how we integrate, how we interact with these artificial intelligence machines and algorithms.

But that brings up a question.

[00:15:22] Machine Speed and Escalation

Ryan Vest: One of the most common things that I see as people discuss AI is they talk about it accelerating things like warfare or decisions to machine-speed. How does this increased speed interact with human cognition and stress, and what kind of risks do we need to worry about, especially when we're talking about nuclear crisis?

Michael Horowitz: I'm sure I wrote something like that. I mean, I published an article on AI and the balance of power for the *Texas National Security Review* back in 2018, which was I think one of the earlier articles, probably, like, on these topics in general. And thank you very much. And the—one of the things you see here is the issue is if you could have algorithmic driven decision-making going, like, faster than humans, which is one of the advantages, then there are clearly some downsides as well.

The question in some ways is whether what we are talking about is tactical or operational and that, at the tactical level, say, if you imagine something like collaborative combat aircraft, but that are, you know, like, fully autonomous, then they're still doing missions that, you know, humans go tell them to do and they can do them faster and more effectively.

And so you're fighting at machine-speed, but you're still, like, under the supervision of a human. Where I think people get really worried is at the operational level. Say, if you had a, you know, the combatant commander for US forces in the Indo-Pacific sort of delegating control of, like, a big conflict against a really large country to an algorithm and, like, letting the algorithm, like, designate, you know, alright, like, which targets, like, which shooters, which sensors, like, that's, I think, the thing that people worry about sometimes

is when the algorithm shifts from being a teammate and a helper to being a —to sort of replacing judgment and, you know, speeding up conflict.

Because the question then is just like, if an accident happens, if something bad happens, like, how does that influence then the prospect for escalation? And to me some of the biggest risks here surround countries that view themselves as having unstable second-strike capabilities.

I think the basic idea is that the more robust, the more secure your second-strike capabilities are, the less likely you are as a military to over rely on AI and delegate, essentially, decision-making within nuclear command and control to artificial intelligence. You might still be at higher risk of accidents for, like, all the reasons that we've been talking about—you know, like automation bias, like those kinds of things— but fundamentally, if you believe in deterrence and you believe in second-strike capabilities, then it doesn't matter how fast somebody is coming at you. You will always have the ability to retaliate.

So, even if somebody's coming at you at machine-speed and somehow, like, destroys some of your military, you still have second-strike capabilities that are robust and able to respond, which in theory means then that nuclear deterrence should hold. The problem is if you are a country with nuclear weapons that doesn't have stable second-strike capabilities, the incentives may change.

Now imagine you are facing an adversary, even a conventionally armed adversary, fighting at machine-speed that you worry now has the ability to decapitate you and decapitate you before, say, you could give a launch order or, you know, do something to try to, like, place your nuclear forces on alert. The risk is that that creates incentives then for nuclear powers that don't have stable second-strike capabilities to engage in all of the worst, most dangerous, like, launch posture and doctrine decisions from the Cold War. Things like launch on warning, pre-delegation, you know, all of that, like, "good stuff," essentially because they're seeking to avoid being decapitated at machine-speed.

And so, this is another way of saying, so much of the writing on this topic is basically about, like, what could happen to the US nuclear enterprise in a world of AI integration. Don't worry about the US in this context. It's been very clear that the US wants nothing to do with AI integration into the, like, deep elements of nuclear command and control. Worry about North Korea. Worry about, you know, countries that we think have a lot less stable second-strike capabilities, because those are the ones that are going to be at much higher risk in an era of AI from machine-speed war increasing the risk of escalation.

[00:19:45] Second Strike and Country Cases

Sheena Chestnut Greitens: So, I was just actually about to ask you if we could put some proper nouns in the conversation and start talking about some countries. How does this explain what we're seeing across the nuclear powers, either established Russia, China, the United States, you know, with large nuclear arsenals and what we think is secure second-strike capability versus some of the newer nuclear powers? How does this variation in perceptions of secure second-strike capability help us understand what we're seeing about how these different countries are responding to AI in their nuclear enterprises?

Michael Horowitz: So, we've—there are a couple of trends I think we've seen in how countries are thinking about AI and nuclear weapons. And with, like, caveat, no idea what North Korea actually thinks. The, like, I mean, Sheena, you would know better than anybody, like tough to get data on this kind of topic.

Sheena Chestnut Greitens: Especially that question.

Michael Horowitz: Like, right, like in particular. And so what we have seen is that I think it's interesting that every country that has looked at this question and published something on it, and that's the US, the UK, France, and China, has basically said, we don't want anything to do with AI-enabled decision-making over nuclear weapons.

So, we saw this first in the US in the 2022 Nuclear Posture Review—shout out to our friend Vipin— and we then, you know, we saw this in a sort of tri-lat NPT statement by the US, UK, and France, and then in the Biden-Xi agreement in, you know, like November 2024, I think. You know, everybody who's like really looked at this closely has said, "Human control over nuclear weapons," you know, "that's the way we want to go." Now, that doesn't mean though that countries might not then integrate AI into—you know, still believe that they would have control in a pre-delegation situation or a launch on warning situation. That doesn't necessarily mean that there's not human control.

Russia's sort of the wild card in this context. Russia has been—the, like, fragments of info we seem to have on Russia suggest that they are similarly skeptical about the value, essentially, of delegating nuclear launch decisions to machines. On the other hand, there's two pieces of contrary information for Russia.

You know, one is of course their historical Dead Hand perimeter system, which was, you know, not an AI system, but a, you know, essentially like a, I mean, like a set of, like, levers and pulleys, basically—whatever, like a radio-based system, like, from the Cold War designed to enable a Soviet nuclear launch, even in a decapitation situation, which, like, fits the logic in some ways of what I'm talking about. So, like, that's piece one. They've already shown interest in this.

Piece two is, if you look at what they've done with the Poseidon system, this uncrewed underwater vehicle that Russia is developing and—there's some public sources suggesting—has, like, at least tested that would be nuclear-armed, where the idea is you could take essentially a uncrewed nuclear-armed submarine and put it, like, off the coast of the United States, say, like, just in case you ever needed it. And, like, I got a lot of, like, operational questions about how that's going to work, to be honest.

But the Russians have clearly shown a little more interest in some ways in delegating. This makes sense in some ways, depending on how you think about the role of regime type here. Like Sheena, you've heard me talk before that the—like, I—when I worry about, like, bad decisions when it comes to AI and military systems in general, but especially nuclear weapons, I worry about autocracies a lot more than I worry about democracies. And the reason is that autocratic countries do not trust their people to begin with. If they did, they would have different kinds of governments.

I mean, look at Xi's consolidation over China's military. If you're an autocratic country, anything you can do to cut pesky, unreliable people out of the, like, chain of command and chain of decision, there's arguably an upside to. And so, to the extent that AI offers the temptation or the reality of letting you skinny down the number of people that you need to rely on prior to making decisions, whether those decisions are, like, using military force against your population, like, or launching nuclear weapons, I think that there are some auto—not that there aren't risks for democracies—but I think that there are some autocratic risks here and, like, there's an autocratic logic of centralization in an era of AI that I worry about a little bit in the nuclear weapons context here, independent of the, like, second-strike argument I was making.

Sheena Chestnut Greitens: Yeah. Well, thanks for getting to the autocracy, authoritarian regime question because you knew I was going to go there at some point.

Michael Horowitz: I mean, how could we not talk about it, given that, like, you are the one interviewing me.

Sheena Chestnut Greitens: Well, I'm glad you covered it because I was going to get there anyway.

[00:24:41] Human Control and Verification

Sheena Chestnut Greitens: Let me then ask a sort of, maybe a little bit of a cynical question, which is, you know, several states, as you mentioned, autocracies and democracies, have emphasized keeping humans in the loop, right? You think about the Biden-Xi agreement and the other countries, I think you said Britain and France as well, have emphasized keeping humans in the loop for nuclear decision-making.

How meaningful do you think those commitments are? I mean, that readout—the Biden-Xi readout—was like a couple of sentences, right? And so is that a credible political assurance, or is this kind of temporary and tactical? Because it's really hard to verify any use of AI in nuclear command and control. I mean, a lot of arms control before depended on some verification. Verification of AI in nuclear command and control—those two things put together seem incredibly hard to come up with any sort of verification regime.

So, are these, sort of, going to be durable norms? Are they tactical? Are they political commitments that could change with leaders or regime change? Like, how do we think about the current landscape of the idea or the commitment, the norm of keeping humans in the loop?

Michael Horowitz: Arms control? I feel like, back in the recesses of my brain, there was, like, a thing called arms control. But since we're, you know, we're recording this right after the end of New Start, and the—I think that this question surrounding arms control and verification when it comes to the nuclear AI nexus is kind of tricky.

I think, first thing I would say is I have been skeptical for years that binding arms control, that verifiable arms control is possible in this arena, for the most part, simply for exactly the reason you said, Sheena, which is that for most countries, the countries aren't going to let you, like, plug into their nuclear systems and check their software to see exactly how autonomous it, you know, it is. You know, think about how hard it was to get people to even get to counting missiles during the Cold War or immediately after the Cold War.

And so, I think that's a really heavy lift to ask states to do, so I am not that optimistic about the possibility of verification in this. And so, if what you need is verification to feel confident, then I think unfortunately states are unlikely to get it. I think the question is the chain of logic that would lead a country to AI-enabled decision-making surrounding nuclear use. And here, actually, I think that this is a place where the incentives of leaders might cut in a way that could make one less worried about the impact on stability. And here's why.

We just talked about the sort of autocratic logic of control over AI and some of the risks that arguably generates if you skinny down the number of people that you have to rely on. But what do leaders like more than anything else? They like control. That's frankly true, whether it's, like, democratic leaders or autocratic leaders. And so, Dan Reiter in his new book, " Untied," " Untying Hands" or "Untied Hands," argues that, in some ways it's like launching a broadside against the spiral model of war, which I'm actually a hundred percent here for. But as part of that, he argues that leaders actually have a lot more control over escalation dynamics than a lot of IR theory tends to assume, because they wish to have the final say, and so always keep off-ramps for themselves.

And if one believes that, and since he was my undergraduate advisor, I feel compelled to say that it's obviously correct—h e was my undergraduate advisor; I do actually think that there's real logic there. And there's an incentive away from the most dangerous kind of decision-making, which would be kind of outsourcing, like, decision-making to an AI surrounding nuclear use. Because if it was credible, if it's not a team, if you're just delegating and delegating, like, you've thrown out the steering wheel. Now the leader's giving up control.

And everything we know about how leaders make decisions in crises suggest that they want control. In part, they want control because they want off ramps, because they don't—nobody wants to end up in a fight if they didn't wish to have it. And so you have kind of like that logic, which should suggest the desire for control, in some ways, like, countervailed by some of what I was saying before, which is the, you know, you are a country with less stable second-strike capabilities in a war against machine-speed and the incentives that creates essentially for, like, relatively dangerous decision-making.

Up to this point, it seems like that logic of control has been more predominant in how countries are thinking about making decisions, which is why we've seen some of the announcements that we have seen surrounding AI and nuclear weapons. But it's going to be really hard to verify in general, and it's going to be especially hard to verify as this era of AI continues and all military systems

have some AI integrated in them in one way or another, frankly, and it's probably not going to be frontier AI all the time.

The question that will then become: Is there, like, some sort of, like, systemic or, like, aggregate risk in, like, a Scott Sagan-y kind of way that maybe is, like, tough to, like, measure or, like, operationalize? Or maybe, like, if you want to know, like, hey, like, why might I be wrong in how I think about this or how, like, the logic in this paper would be wrong?

An argument against the logic of this paper for why you should be more worried would be if you think that there are going to be all these little integrations of AI, they're going to be relatively hard to track. They're all going to be, like, pretty good, not perfect, and that increases, sort of, the interaction between those, like, increases risk in some ways that we can, like—that's hard to articulate frankly, like, let alone measure. Like, I would say that, like, that's not a— if you can't articulate what the risk is, like, or what are we even talking about? That's an argument one could, I think, certainly make.

[00:30:35] Policy Takeaways and Psychology

Ryan Vest: We seem to live in a sea of pessimism around AI and the future of AI, especially in nuclear arms control. But I'm really impressed that as we've talked today, you have this very optimistic tone, this very optimistic look at the future.

But with that in mind, you know, because I think you've got a unique perspective here—in reading through the article, I really loved a lot of the things you had to say—what do policymakers need to take away from this? What do they need to be thinking about on how we should implement AI and what we should be concerned about with our enemies implementing AI in the future?

Michael Horowitz: No, that's a really good question. I mean, I tend to be a little more sanguine on escalation risk, I think, in some cases than others. I think it's in part because of a project I did, I worked on, about a decade ago on geopolitical forecasting in quantifying geopolitical risk, which made me think that there's more— there can be, sort of, like, escalation risk inflation t hat happens out there. And so when I hear lots of people saying, like, "OMG, the escalation risk," it, like, sometimes takes me like, "Is there? Is it? How bad is it?"

A thing that one could worry about, like, honestly, is suppose you believe that the argument that I make in the paper or that I make here about second-strike capabilities and AI integration risk and basically, like, bad decision-making surrounding AI. There's a story that one could tell that, frankly, only the United States has reliable second-strike capabilities. We could say, like, "Sure, like, North Korea doesn't, like, sure, Mike, that's an easy case for your argument." But, like, Russia and China have worried for decades about the possibility of American-driven decapitation, especially with advancing American nuclear capabilities.

China's nuclear buildup is multifaceted, but the—I mean, like, empirically it's multifaceted—but the logic behind it is also multifaceted. But, frankly, the desire to increase their reliability of their nuclear capabilities in a world where the US could, you know, launch a first strike is, I think, something that's certainly, you know, like, seems like it's, you know, been on their minds.

And so, in that way, if you imagine a world where everybody believes that they don't have secure second-strike capabilities relative to somebody else all the way up to the United States, then there frankly could be a logic if AI-enabled conventional capabilities get strong enough and accurate enough that essentially everybody but the United States then might feel some—maybe France and the UK, depending on how like allies and partners things are working then— might have an incentive then to engage in some of that pre-delegation or launch on warning or sort of riskier decision-making because they fear machine-speed decapitation.

But this gets back to something that we talked about before, which is that this is all a play on psychology in some ways at the end of the day. Like, we're talking about artificial intelligence, but what are the factors that, to me at least, are most salient in thinking about escalation risk? It's not the technology in and of itself, because militaries have testing and evaluation processes and validation verification processes and all sorts of things designed to ensure that when systems are fielded, they actually do the things that they're supposed to do and don't do the things that they're not supposed to do. So, like, let's presume at like a basic level that these systems work, because, frankly, if they don't, then those systems won't get fielded.

The issue is that it's how humans interact with them, like, what training and education looks like, and how policymakers think about escalation risk. For example, it is easy to imagine senior decision-makers in the US government or any government either being, frankly, algorithm averse because they got to the positions they got to because of their expertise, not by relying on a computer,

and so disregarding advice from an algorithm when they should accept it, or becoming infatuated with whizbang technology and falling prey to automation bias.

The reality will probably depend on the situation and the person, which is not, like, comforting from a, like, a broad international relations theory perspective, but I think reflects the messiness of the way that policymaking would work. And that's especially true if what we are talking about is a low probability, high impact event like the use of a nuclear weapon, which is not like a regular Tuesday for anybody, essentially.

That's exactly when, in some ways, you would imagine people to, like, gut check to, like, how they think. And how that then interacts with these AI systems is tough to systematically anticipate.

Sheena Chestnut Greitens: So I think this is one of the really interesting things about the collection of essays that your article appears in, because some of them probe the sources of excessive techno-optimism, and then others point out, you know, as your essay does that things like, you know, military conservatism to these testing regimes you just mentioned, bureaucratic caution and operating procedures could kind of constrain the adoption of really risky applications of AI.

But I'm still struck by this point that confidence is not linear over time.

[00:35:31] AI Literacy and Moving Faster

Sheena Chestnut Greitens: And I'm wondering, just as kind of a, maybe, a final question— you sit at Perry World House at UPenn, you're working with, thinking about, students and policymakers that are going into or advancing through careers in national security. And you talk about in the article, the idea that at different points in the cycle, there's—this trust gap is bigger or smaller and can actually go in a different direction. So you can have underconfidence or overconfidence in AI depending on where you are in that time curve.

So is there a way for universities or people who train policymakers in a whole bunch of contexts or just people who might be called on to use AI and want to figure out how to do it better rather than worse—how should they think about—is there something they can do or that we can do as a national security enterprise to better calibrate peoples' assessments of what AI can and can't do and how accurate it is to use it well?

Michael Horowitz: That's a great question. There are a bunch of different levels of this. I think, like, to me, this is an argument for AI literacy for everybody. This is not "everybody needs to be a coder." It's not, you know, "everybody needs to be able to train the next generation of frontier models," but everybody being familiar with how AI works and how various AI systems work. And there are a growing number of training education programs, frankly, that are designed to do just that. So, like, I mean, that I think is the, like, bumper sticker kind of takeaway on that.

Like, one thing that's encouraging in this context is if you look at what the US military service academies are doing, they've all introduced, in different ways—which is fine, frankly, like, we're early in the era of this and, like, this should be like a let a thousand flowers bloom world in some ways for training and education on responsible AI, or on using AI responsibly is probably a better way to say that than like the responsible AI bumper sticker—but if you look at the Naval Academy or West Point, you look at the Air Force Academy, they all have introduced AI curricular elements, both, you know, within, you know, where you would expect them in the classroom, like computer science and engineering, but also in how they're, like, teaching about the future of war, how they're, like, teaching cadets to, like, think in the context of leading on the battlefield, whichever domain that battlefield might be in.

And, you know, to the extent that we think that more knowledge in this space—from the experimental data—that more knowledge in this space tends to be better because more knowledge tends to lead to more understanding of the limits of algorithms, and thus, like, relatively better calibration, then you would, then, like, that kind of training is a great first step and that it can, in some ways, then we end up almost in a transition period, where, you know, you have in some ways, like, your junior officers and probably even junior policymakers that are increasingly going to be, like, AI natives and, frankly, understanding this because, like, they, like, come up, like, using AI systems in high school and, like, understand that, like, how fake a bunch of the things that they can get out of it could be. Like, they might be, like, way better socialized on that than maybe like us.

And then the question becomes, especially for the next like decade or so, as you have senior decision-makers coming into positions of authority that maybe don't have that kind of background: How is it that you do training with them to get them up to speed? Because they're the ones in some ways that shoulder the biggest cognitive load in making decisions in general. And so their decisions about when and how to use AI will sort of, like, have the highest, like, leverage.

But, like, to me, in some ways this is, like, all about humility, about like our own knowledge and about what the systems can do. But to wrap with something on a slightly different axis, all of that is true. I do not think that is an excuse to go slow. I think that when we look at military innovation adoption in the United States, both, like, broadly, but that certainly in this era, there's way more going too slowly rather than too quickly. And people are quick to point to experiments or prototypes or like random things from DARPA as, like, you know, evidence of some, like, terrible thing that might happen, when in reality that's the system working. And, frankly, for the uses of AI that are proven to be effective, whether in the nuclear realm or not, the Pentagon has been slow to scale and it's frankly time to hit the accelerator on some of those things.

Sheena Chestnut Greitens: That's a really good point too, that there's a risk, not just to adopting AI, but also not adopting it or being too slow to adopt it in a world where other people, other actors are.

[00:40:10] Closing and Credits

Sheena Chestnut Greitens: This has been a really interesting discussion, Mike. Thanks so much for joining us today.

Michael Horowitz: Thank you for having me.

Sheena Chestnut Greitens: Thanks for joining us for *Horns of a Dilemma* the podcast of the *Texas National Security Review*. Our guest today has been Michael Horowitz, author of the article, "Artificial Intelligence and the Future of Strategic Stability," which as always can be accessed for free on our website, TNSR.org, with the rest of the special issue in which his article appears. If you enjoyed this episode, please be sure to subscribe and leave a review wherever you listen, and you can find all of our work at TNSR.org. Today's episode was produced by *TNSR* Digital and Technical Manager Jordan Morning and made possible by The University of Texas System. This is Sheena Chestnut Greitens and Ryan Vest.

Thanks again for listening.