Center for Strategic and International Studies

TRANSCRIPT

Event

"The DARPA Perspective on AI and Autonomy at the DOD"

DATE

Wednesday, March 27, 2024 at 10:00 a.m. ET

FEATURING

Matt Turek

Deputy Director for the Information Innovation Office (I2O), DARPA

CSIS EXPERTS

Gregory C. Allen

Director, Wadhwani Center for AI and Advanced Technologies, CSIS

Transcript By
Superior Transcriptions LLC
www.superiortranscriptions.com

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES

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

Today, I'm joined by Dr. Matt Turek, the deputy director of the Information Innovation Directorate at DARPA, the Defense Advanced Research Projects Agency. DARPA is a legendary organization in the history of so many technologies, not least of which are AI and autonomy. And that's the focus of our conversation today.

Dr. Matt Turek, thank you so much for coming to CSIS.

Dr. Matt Turek: Thanks. Really looking forward to the conversation.

Mr. Allen: Before we get into the meat of what DARPA's up to in AI and autonomy, I wanted to get a little bit about your own background and sort of how you got into the field of AI, and how you got into the field of military AI and autonomy. So how did you – how did you come work in this field?

Dr. Turek: Yeah. That's a great question. I started my career – I was lucky, actually, at GE Medical Systems, as part of that as an engineering program. So entry-level rotational engineering program. And really started with a great cohort of people on that program. That got me broad experience across GE Medical Systems. This was the mid-'90s. And so, you know, that was times when there was early interest in, you know, having automated algorithms as second readers on medical image analysis.

Mr. Allen: And GE is very well known for all kinds of medical scanners, like CAT scans, MRIs. They're big in that kind of sensor technology.

Dr. Turek: Right, yeah. And that was the business at the time. And, of course, GE has morphed quite a bit since then. And I spent some time at GE Medical Systems. And then I had a really excellent opportunity to move to the GE Global Research Center and work in a research environment that both serves industry, but also interacted with government. And then from there, I left actually and went and got a Ph.D. So that gave me that academic background, provided some of that academic rigor. And, you know, obviously that has been really useful.

And then after that, I joined a small business and was there for about 10 years or so. Helped run a computer vision team. And that's really the time in which I started working more on AI, particularly computer vision, in military-relevant domains. And a lot of that was funding from agencies like DARPA, but also working with Air Force Research Labs, and NGA, and others. And what ultimately became really attractive for me was just knowing that I was serving this broader mission and sort of feeling that tangibly. And, you

know, that's something that has made my time at DARPA, after I left that small business in 2018. You know, that has really been sort of the fuel for passion of the work at DARPA, is that ability to be in – to do work in service to the war fighter.

Mr. Allen: And so you've been a part of the computer vision AI revolution both in the handcrafted algorithms era all the way to the sort of modern machine learning and neural networks part of the story – and been serving your country during that transformation, which is really exciting. So DARPA is, of course, well known all around the world as a legendary technology organization. But there's so much about what's going on in AI and autonomy that DARPA has been central to, both in the past and in the present. So how do you sort of sum up what DARPA is up to in the AI and autonomy field today?

Dr. Turek: Yeah. That's a great question. I mean, I guess I'll just start as a – with a reminder that, like, DARPA started in 1958. We've been investing in AI probably since the early '60s. You know, so going back to just a few years after the coining of the term AI. I2O specifically, the Information Innovation Office within DARPA, has really four key thrust areas. So, proficient artificial intelligence is one of them. Confidence in the information domain. So that might be tools that help us understand things like manipulated media. Building secure and resilient systems. And then tools in cyber, both defensive and offensive.

And there's a lot of synergies across those thrust areas. So, you know, we have efforts that are blending both, you know, advancing AI and advancing the state of capability in cyber. There's interactions between AI – again, sort of core AI algorithms and the development of tools that might help us understand things like manipulated media. That's within I2O. There's five other technical offices. And I think it's worth saying that, you know, AI and autonomy is really being used broadly across the agency now.

Probably something like 70 percent of our programs have some type of AI, machine learning, autonomy associated with it. So there is really broad penetration across the agency. So it's really difficult to sum up, you know, what the agency as a whole is up to, but from an I2O perspective we're really looking to try and advance, you know, how do we get to highly trustworthy AI – AI that we can bet our lives on – and that not be a foolish thing to do.

Mr. Allen: That's an incredible line, by the way – AI that we can bet our lives on, and that not be a foolish thing to do. I'll have to remember that one.
I want to ask a little bit about how DARPA works. Because on the DARPA resume is incredible things like stealth technology, like the invention of the internet. But DARPA has an incredibly diverse project portfolio. And it also has an incredibly diverse project management toolkit. So could you just

explain what are the sorts of different types of DARPA projects, how they accomplish their goals?

Dr. Turek: Yeah. I mean, again, DARPA's core mission when we were founded was to prevent strategic surprise. And we also think about creating strategic surprise. Again, the founding of the agency was tied to Sputnik. It was something that created –

Mr. Allen: Which was one heck of a surprise, yeah.

Dr. Turek: Created a strategic surprise for the U.S. It wasn't necessarily a surprise that they were launching a satellite. The details that were actually what was surprising. What was the size of the payload? And the implications of that – the size of the payload – meant, well, they perhaps could put a nuclear weapon in orbit on an ICBM.

Mr. Allen: If they could do this, then they can do that.

Dr. Turek: Exactly. And so that, for sure, created a strategic surprise. And again, that, just a handful of months later, led to the creation of DARPA, with, like, a page-and-a-half memo. So just think about, you know, standing up a multibillion-dollar agency now with a with a page-and-a-half memo. But, you know, back to your question about, you know, how do we think about investments and the project portfolio, and how do we manage them? So, you know, DARPA, in service to that preventing or creating strategic surprise, really looks to be disruptive. You know, can we disrupt adversary capabilities by coming up with a new defensive capability? Can we provide a new strategic

capability for the U.S. that disrupts what adversaries are able to do? And you mentioned things like stealth and GPS. And those are some of the classical examples.

So how do we get to that level of disruption? You know, one of the things that really starts with, actually, is just hiring great program managers. And we always need to hire program managers. Everyone is on a clock at DARPA. So that forces turnover. And so, you know, ideas are very bottom-up-driven within DARPA. And then, again, with that lens towards disruption, it might be, do we just need to make an investment to help instantiate a research community in a particular space that's necessary to the DOD? Do we need to build a transformative capability for our war fighter and get it in their hands as quickly as possible? Those are sort of two different poles on a continuum of technology. And we really look at those particular endpoints now to help shape how we think, you know, about the investment process.

Mr. Allen: And so in the former case, where you're trying to create a research community, this might just be something, like, we think this is an interesting area and we wish people were exploring it. So we're going to start making grants for people who want to do that. Which is vaguely analogous to the way that the National Science Foundation might do its work. But then on the other hand, as you said, DARPA can actually conceptualize an idea, run it all the way to not just the development of the prototype, but maybe even a version that that war fighters get access to. So there's a really diverse set of project types.

Dr. Turek: Yeah. And on that, you know, the comment about the instantiating a research community, you know, it's unlikely that we're going to say, well, you know, we're going to behave like NSF and we're just going to sort of cede some money. We're going to approach it with a particular purpose. You know, tracing back to a perceived gap in our technical capabilities or a perceived gap in our understanding of technical capabilities, we really feel like more research in this area is needed. And so we can help create that research by forming a program with particular problems and then funding researchers to carry that out.

That might be done, again, with a plan that, hey, we need to help create a research community in a particular space, or maybe help balance research communities. Again, you know, this traces back to the fact that DOD needs, and industry needs, or, you know, academia needs may be different. So, if we make those sorts of investments, they will be very intentional in order to, again, help create a technology base that can be transformational for the war fighter. And then, yes, that other endpoint, you know, there – maybe there is a near-term pressing need that no one else can meet. Or, we have a transformative idea that, you know, would be highly beneficial if we could rapidly get that in the hands of the war fighter. And so, yeah. That helps – those two endpoints on that spectrum really help inform our thinking about investments.

Mr. Allen: So in just a moment we're going to talk about examples of programs that DARPA is currently running in AI and autonomy in both of those types of categories. But before we do that, I wanted to ask sort of how DARPA fits into the DOD picture. Because while DARPA, you know, has a long, storied history in AI, as you mentioned, there's other organizations that have been created around technology adoption, technology innovation, such as the Defense Innovation Unit, DIU, which is now a direct report to the secretary of defense. There's also the Office of the Chief Digital and AI Officer, the CDAO organization, which reports to the DepSecDef. And so I'm curious, you know, sort of how does DARPA fit into the DOD portfolio of organizations working on AI? And I guess the other part that we didn't mention is the service labs and the service programs of record. So where does DARPA fit into this story?

Dr. Turek: Yeah. And those are – that's a great question. And those are all partners that that we work with throughout sort of the spectrum of technology.

Mr. Allen: I guess I should – I should acknowledge here that the CDAO's predecessor organization, the Joint AI Center, where I worked, was actually a customer of your work specifically related to the to the MediFor program on deepfake detection.

Dr. Turek: Yeah, that – I mean, that was – that's one example of the sort of collaboration, and it's actually deepened with CDAO across – in particular – across multiple programs. But let me start by giving that – sort of giving that broad perspective, and then maybe I can give you a couple examples of places where there's – where there's collaboration.

So, again, DARPA's core mission, prevent and create strategic surprise. So the implication there is that we're looking over the horizon for transformative capabilities. So in some sense, we are very early in the research pipeline typically. Products that come out of those research programs could go a couple places. They can stay within the DOD and then transitioning them to CDAO, for instance, might enable broad transition across the entirety of the DOD. You know, I'm actually happy that the JAIC was stood up, that CDAO was there, because I think having an organization that can provide some shared resources and capabilities across the department, can be a resource or a place where people can go look for help or tools or capabilities, I think that's really useful.

And, from a DARPA perspective, it gives us a natural transition partner. So, yes, on our Media Forensics program, we transitioned algorithms over to the Joint AI center for assessment and to just demonstrate across the force. We continue to do that with other programs, like our Guaranteeing AI Robustness Against Deception program. So that is a program that's focused on building defenses against adversarial attacks on AI systems. So whether that is physically realizable attacks or noise patterns that are added to AI systems, the GARD program has built state-of-the-art defenses against those. And some of those tools and capabilities have been provided to CDAO.

Mr. Allen: Can you just talk a minute – because I think a lot of our audience will have heard of adversarial AI, but perhaps not all. So what is the sort of problem you're trying to solve here in the GARD program, specifically?

Dr. Turek: Yeah. So one of the things that's – well, I guess, two starting points for AI systems. So AI systems are made out of software, obviously, right? So they inherit all the cyber vulnerabilities. And those are important class of vulnerabilities, but not what I'm talking about here. There are sort of unique classes of vulnerabilities for AI or autonomous systems where you can do things like insert noise patterns into sensor data that might cause an AI

system to misclassify. So you can essentially, by adding noise to an image or a sensor, perhaps break a downstream machine learning algorithm.

You can also, with knowledge of that algorithm, sometimes create physically realizable attacks. So you can generate very purposefully a particular sticker that you could put on a physical object, that when the data is collected, when that object shows up in an image, that that particular what's called adversarial patch makes it so that the machine learning algorithm might not recognize that object exists or might misclassify that tank as a school bus. So those are sort of classical examples. You know, there's other classical examples of placing a sticker on a stop sign and causing a machine learning system to misclassify that as a speed limit sign, for instance.

Mr. Allen: Yeah. So what you're – what you're getting at here is that every AI system is sort of a combination of traditional software and machine learning software. And you can hack those systems either by hacking the traditional software, but what you're getting at is there's this entire new category of hacks which is often called adversarial AI. And you're trying to think about how do DOD systems have safeguards embedded so that they're not vulnerable to this sort of category of attacks.

Dr. Turek: Yeah, exactly. And not only are we thinking about it, we have created new algorithms. Some of those actually are in partnership both with the research teams that we're funding but with researchers at Google, and then created open-source tools that we can provide back to the broader community, so that we can really raise defenses broadly in AI and machine learning. But those tools also provided to CDAO, and then they can be customized for DOD use cases and needs. And so there's, you know, a multipronged transition strategy.

So anyways, that's a concrete example of, you know, how we might work with CDAO. On the Defense Innovation Unit side, you know, some of the foundational research investments from DARPA might get commercialized. They might become commercial industries. And that provides an opportunity for folks like DIU, that might take the best of breed of what's available commercially and bring that rapidly into the – back into the department.

Mr. Allen: Right, because DIU sort of sees themselves as the front door to DOD for this sort of commercial technology sector. But that commercial technology sector might have been harvesting investments that DARPA made a while ago.

Dr. Turek: Yeah. And it actually turns out that sometimes the most efficient way to get the technology into the DOD and broadly dispersed is to go through that commercial route. And that avoids some of the traditional, you know, operations and maintenance and sustainment funding issues, where you

actually have a commercial entity who has a business model that includes supporting the DOD, but that also might include supporting, you know, the broader technology base within the U.S. And particularly in the spaces I2O works in – you know, information domain, AI, cyber – you know, it's not just U.S. government systems that need to be protected. It's, you know, the technology base, critical infrastructure broadly speaking across the U.S. Those are also attack surfaces for an adversary.

Mr. Allen: Great. And so before we go into sort of program by program, which I think is going to be fascinating, I do want to get your sense of AI writ large. Sort of what is this moment that we're currently in? Because the machine learning revolution in it sort of modern form, which really took off in 2012, has been underway for more than a decade. And now it seems like we have this additional revolution. And some folks are talking about human level AI across a broad range of categories in the not-too-distant future. You've been in this field, deep in the weeds, deep in the research community, and now a leader, you know, in the research community. Where do you see the sort of current moment where we are in AI and autonomy?

Dr. Turek: Yeah. Yeah, that's a good framing in terms of, you know, really the explosion that happened around 2012 or so. And just to put a point on it, you know, that was really the use of AlexNet and a deep learning approach on a computer vision benchmark that really caught the research community's attention. Like there was just a significant step change in performance. And what you saw in the research community in terms of both academia and industry – as evidenced at conferences like CVPR, which is Computer Vision Pattern Recognition, one of the top AI conferences in the world – was just this massive shift over a relatively short time where everybody was leaning in to using these sorts of deep learning approaches. And then – you know, so that's 2012.

About 2014 or so, Ian Goodfellow comes – and others – come up with generative adversarial networks. And for me, that's – you know, that's a similar sort of explosion point in what we now call generative AI, right? And so these generative adversarial networks was really – a really interesting

insight between instead of trying to train to a particular objective function, I'm going to compete a deep neural network that can generate a piece of data with another deep neural network that is going to try and detect whether that piece of media was synthetically generated, or whether it's real or not. And that really, I think, helped sort of further the explosion of deep learning.

And then we started to see folks using – you know, moving from computer vision into natural language processing, and using things like transformer models to do token prediction. So, like, what is the next word or what is the next fraction of word? Can I predict that? And that is – that really basic

sounding capability is what really underlies things like ChatGPT and the state-of-the-art in large language models. And so that is what has everybody's attention these days.

And, you know, what is explicit for some but maybe implicit for folks that are not embedded in the community, is this notion of a scaling hypothesis. And so that is really a hypothesis that if we make larger and larger models with more and more parameters, and we feed them with more and more data, that is going to get us to more and more intelligent systems. And the data – there are actual scaling laws that have been experimentally derived. So you can see that there are actually trend lines. And those scaling laws are all based on what is my accuracy in predicting the next – the next token? And the contention is that in order to do that prediction better and better, I have to actually build an underlying model of the world. And that will get us to intelligent systems.

For me, I still feel like that's a hypothesis. You know, I don't know what the ceiling is on that – on that capability. And so one of the things I've said before and I'll say here is, like, this is the time of my career where I actually have the most uncertainty about what is the right technical approach, what is the right technical thing to do. And I feel like having some technical humility is a really useful approach. You know, folks from the AI community might think about that as having a more probabilistic model. If you make a hard decision, then then things can break down. So carrying that uncertainty through your thought process, I think, is –

Mr. Allen: So I think this is – this is super interesting. So the scaling hypothesis – I'm oversimplifying here, right – but it basically says: If you take the existing set of algorithms, the sort of same algorithms that are already powering ChatGPT and its equivalents elsewhere in commercial industry, and you simply feed them more data for training, and more computational power for, you know, running those that training approach, then the performance of the system will get better and better and better. And the question is, is that true? Or is this going to plateau at some point sort of short of human intelligence?

And this is a debate among AI researchers around the world. And I think it's quite interesting that you do not see enough evidence to discount this hypothesis at this current moment, right? It could be wrong. But it also couldn't be right. And we should operate with that understanding, that it may be the case.

And then the other flip side of that is algorithms have made a lot of progress over the past 10 years, and that doesn't show any signs of plateauing anytime soon. And so we are currently in a world where AI is really impressive but. correct me if I'm wrong here, you're not seeing anything to discount the possibility that we could be dealing with systems that are not

just two times better than the current state-of-the-art, but 10 times better, 100 times better, 1,000 times better, you know, in a matter of years or decades. Is that – is that fair?

Dr. Turek: Well, you know, again, I think having the uncertainty is important here. So do I know what the ceiling is for the current approaches? No, I do not. Do I think that just – do I think that
we will – that these approaches guarantee that we will build that underlying model sufficiently to get to something like human level intelligence, broadly speaking? I'm skeptical about that.

Mr. Allen: So your sort of hunch, I guess is the way to say it, is that we do need architectural improvements, we do need algorithmic improvements?

Dr. Turek: Yeah. No, I think that's going to be critical. And I think particularly, coming back to DOD needs, you know, how are DOD needs different than industry, right? Well, some of it revolves around our access to data and compute, actually. So you might think, well, like, you know, DOD should have massive amounts of data. Well, state-of-the-art AI systems are essentially being trained on all the data on the internet. So if you look at, you know, U.S. government data holdings in satellite imagery, for instance, you know –

Mr. Allen: Mostly not on the internet. (Laughs.)

Dr. Turek: Well, also, mostly not on the internet. But, you know, all the information that humanity has produced and is on the internet is a pretty high bar for being able to train state-of-the-art AI systems. So in some sense, actually, you know, data is a challenge on the U.S. government side.

I think, also the criticality of the decisions and the sorts of scenarios in which we might want to ultimately use AI and autonomous systems are different from industry. So, you know, industry revolves around trying to find quick business opportunities. What is my business case? How do I service a broad customer base? How do I get that customer base as quickly as possible? And, you know, those are all valid needs to address. In fact, from a national security standpoint, like, we want to have a robust commercial technical base. But that's very different from the DOD, where we might not be able to pick and choose as much where we use AI capabilities. We may be pressed by adversaries, and that might shape – or, you know, the – you know, so that's one thing that's different.

Also, just the criticality of decision-making, right? Like, there are places where, yes, summarization tools, things like LLMs, could certainly help automate processes, you know, particularly automate bureaucratic government processes. But that's a far cry from making a life or death decision, or even a life or death recommendation that a human then needs to

resolve and say, yes, am I going to – am I going to go forward with that decision point? You know, I don't think industry is there for many cases. And, you know, again, from a business perspective, like, that's not the right – that's not the right place to start. So I think that is another fundamental difference between how industry is approaching AI and how the DOD needs to think about it.

Mr. Allen: So I think that's a great transition into what's going on, because you're obviously a very insightful observer of the field of AI and autonomy. But you're not really an observer. You're an actor in this space. You're trying to shape, you know, the trajectory of research. And I2O has a pretty impressive portfolio across this. I want to start with the program that you and I have a little bit of personal history around, which is related to what's commonly known as deepfakes and the detecting of synthetic media. And you have two programs here that have done some really interesting work. So could you talk about MediFor and SemaFor?

Dr. Turek: Sure. Yeah, so MediFor is the Media Forensics program. That was started by a previous program manager, Dave Doermann. And I think we really – actually, the media forensics community owes him a debt of gratitude for sort of foreseeing that this problem was going to exist. So he started building the program in 2015. It kicked off in 2016. Ran for –

Mr. Allen: So only one year after the Ian Goodfellow GANs insight, he was already on the – that's fabulous.

Dr. Turek: Yeah. Well, and some of the motivation was actually tied to just the capabilities in Photoshop, right?

Mr. Allen: Oh, sure. Sure, sure.

Dr. Turek: So it was not just around generative AI.

Mr. Allen: Which North Korea has historically actually used Photoshop to put out fake imagery. So, of course, we needed that.

Dr. Turek: Right. Right. And so that program really focused on images and video. And simply, you know, could we produce algorithms that would have a quantitative measure of – that would create a quantitative measure of integrity for a media asset? So just demonstrating that you could actually quantify the problem. And that quantitative measure itself means that you could automate processes, like prioritization at scale, or filtering. And so that program ran from 2016 to 2020. I inherited it in the summer of 2018. And, you know, that was a great entry point for me at DARPA.

And then the follow-on to that was our Semantic Forensics program, which I designed and started. And that program kicked off in 2020. I handed it over to another program manager, Wil Corvey, when I took on the office leadership role in 2022. But that program is focused not just on detection, but also attribution. So does media come from where it claims it came from? Characterization, was media generated or manipulated for malicious purposes? Super difficult to define. That's probably the hardest problem on that program. You know, there has been some progress in terms of trying to develop a taxonomy of how you might think about malicious uses of the – of the technology.

SemaFor is winding down later this year. And so one of – you know, I mentioned the commercialization path in the context of DIU. One of the things that semaphore is doing is open sourcing some of the algorithms as proof of principle, as – you can think about them as reference implementations that broader commercial community could use to help bootstrap commercial capabilities. Because, you know, it's not just U.S. government that needs to have these capabilities. The attack surface is broad. And, ultimately, you know, a can't be U.S. government that is the sole funder of research and defenses in this space. We really need to create a commercial community. So we put some open-source capabilities out there to help incentivize that commercial development around the MediFor and SemaFor work.

Mr. Allen: And just thinking about the national security logic underpinning a program like this. United States is a democracy. And the quality of democratic debate really depends upon standards of truth. And we have been living in this lovely island of history where the tools for recording media and authenticating media had been superior to the tools for forging media. And that's been true basically since the invention of the camera in the 1800s.

But now we're entering this new era – and we've already sort of been in it for a bit now – where the synthetic media generation tools are really catching up to the authentication tools. And that's a real challenge for elections. That's a challenge for determining war crimes. You know, anytime anything happens in Ukraine there's obviously video that's captured. And we want to know whether or not that open-source stuff actually happened, where it says it happened, under the circumstances in which it

depicts, et cetera. So the United States, it seems to me, has a real interest in being able to authenticate media.

What I want to ask you is, where do you see this headed, right? It seems right now that the authenticators have an edge over the forgers. But it's not the same edge that it used to be. You know, 20 years ago my eyes were most of the time enough to determine whether or not an image was fake, even if

Hollywood spent $100 million on computer graphics. You know, today my eyes are often not good enough. And we need some fancy technology, like what the MediFor and SemaFor program are developing.

Do you think it's likely to be the case that authentication technologies are going to continue outpacing the media generation? Or what do you expect to take place over the next year, decade, et cetera?

Dr. Turek: Yeah, certainly the generation capabilities are becoming much more compelling. They're becoming much more ubiquitous. I think we're going to – we should expect to see them used at speed and scale, maybe for mis- and disinformation, maybe for targeted, large-scale, personalized phishing attacks, for instance. There's already been uses of them in financial fraud. So, again, just more evidence that, you know, the attack surface is broad here.

Where we ultimately land, again, I think this is a place that's difficult to say. Part of the reason why we designed the program the way we did was it could be that generative AI becomes ubiquitous, and then detecting –

Mr. Allen: It certainly seems to be headed right away, yeah.

Dr. Turek: Right. Detecting whether something is generative AI or not isn't as useful. But if you can authenticate where media comes from, well, that's useful, right? So if I can still, you know, attribute media back to a particular development tool or back to an organization or an agency, that is very useful, and provide supporting evidence for credibility. And then, furthermore, if I can automatically assess, like, you know, what might be the intent behind, you know, how media was created and designed, and how it's presented to the user, that also helps provide some additional information beyond real or synthetic. So I think the questions become more difficult. They become more nuanced. I think the role of tools is going to remain important.

That's why I think we want to help create commercial industry in this space, because, again, you know, you used examples from politics and national decision-making. But, you know, insurance companies, online commerce, the scientific process –

Mr. Allen: The basic functioning of the economy and society, yeah.

Dr. Turek: Yeah. I mean, those are critical to national security, but also just to our quality of life. And so I think there are real opportunities here to create commercial industry.

Mr. Allen: So this is an incredible program, now coming to its conclusion via this transition. And I do think it's an incredibly interesting strategic decision to open source these tools, really making a bet that truth and the United States'

national interest are sort of aligned naturally, is a very interesting strategic decision. Not every country would make the same conclusion. (Laughs.) But that's not the only thing that your team, your organization is involved with in generative AI. So can you talk a little bit about the rest of the generative AI portfolio, besides the authentication part?

Dr. Turek: Yeah. I mean, one of the unique things that we're doing actually is the AI cyber challenge. And so that was released – that is literally going to be a competition to try and use generative AI technologies, like large language models, to automatically find and hopefully fix vulnerabilities in open-source software, particularly open-source software that underlies critical infrastructure.

Mr. Allen: So there's a bit of history with this cyber grand challenge, right? I believe the last cyber grand challenge, correct me if I'm wrong, was 2016, something like that?

Dr. Turek: That sounds – I don't know that I know the date for certain. But that is the right – the right timeframe.

Mr. Allen: And that was – that was an impressive demonstration of autonomous cyber capabilities. But what I think is interesting is there was no machine learning among any of the teams that were running those autonomous cyber systems. This time around, with the cyber grand challenge, I think everybody's using machine learning to some greater or lesser extent.

Dr. Turek: Yeah. I mean, you know, one of the unique things about the design of the AI cyber challenge is the partnership with state-of-the-art LLM providers, like Google and Microsoft, and OpenAI, and Anthropic, that are actually providing –

Mr. Allen: They're all participating in a DARPA program.

Dr. Turek: Right. And they're all providing access to state-of-the-art models. And then the competition is set up as a prize competition. So there are millions of dollars in prizes to try and incentivize as broad a community as possible to engage on this problem. You know, we'll see what we – what solutions look like. But, you know, one of the things that we speculate about what compelling solutions might look like, you know, leveraging those large language models but also leveraging more – you know, earlier approaches to AI that are more symbolic based, in terms of cyber reasoning systems, because software –

Mr. Allen: Still useful.

Dr. Turek: Still very useful. And software, in some sense, is naturally about manipulating symbols. You know, how software is written, that's how humans think about it, that's how the code is written. And so, yes, you can derive statistical patterns from them. But there's also that sort of symbolic – that sort of natural symbolic information that you can exploit. And so, you know, again, we'll see what the competition results look like and what the approaches look like. But, you know, one compelling approach might be to leverage cyber reasoning systems that are more symbolic with these compelling statistical models in the context of large language models.

Mr. Allen: So these systems might be sort of hybrid approaches, taking advantage of the more traditional approaches to AI using input/output rules-based systems and, as you said, symbolic logical reasoning. But then also mixed together with the capabilities of new generative AI systems. I do think it's so interesting that one of the languages that large language models are so good at are all the computer programming languages. And that seems to be such a natural fit for cyber. I think the other natural fit for cyber is that modern machine learning systems are all incredibly data hungry. And in the cyber domain, generating data can be done through simulation and digital means.

You know, if you want to create data about Moon launches, you have to launch rockets to the moon. It's very expensive and complicated. But if you want to collect data about network, you know,

intrusions, you can just go run those network intrusion simulations, and generate useful data. So I think what's also very interesting about what you said is this partnership that you have with sort of the leading large language model developers, many of the relevant companies. What's that partnership like? What are they getting out of it? What are they providing? What is DARPA getting out of it?

Dr. Turek: Yeah. I mean, this is really a credit to Perri Adams, who is the program manager that designed the program. And, you know, you'll sort of hear throughout my comments today about the importance of that role of program managers, and, you know, something we're always on the lookout for, just to put a shout out.

Mr. Allen: I mean, it's one of the most desirable jobs in the entire defense ecosystem, and a lot of legendary people at various points in their career have been DARPA program managers.

Dr. Turek: Yeah. I mean, I think it's a really unique opportunity to transform a research community. But, you know, Perri had a lot of insight into this problem and, you know, leveraged connections to start the conversations with those – with the providers of those sorts of models. So, you know, from our

perspective, this provides the DARPA performer base – again, whoever decides to sign up for this challenge – access to state-of-the-art capabilities.

What the companies get is also access to understand, like, oh, that's an interesting use case for my model. Maybe that's something that I – that I didn't think of. And, you know, so, again, we'll see how the competition plays out. But there may be commercial opportunities to build these sorts of defensive systems that can find and fix vulnerabilities. Certainly, some of those large language model companies might ultimately see that as an interesting business model, or partnering with researchers or companies that are working on the program. So I think the benefits for them is just to understand a potentially compelling application area for these large language models that they built.

Mr. Allen: And could you help us understand a little bit your sense of what the future looks like in this domain as well? You know, we've talked about why cyber and AI capabilities are sort of naturally a good fit. And many cyber capabilities are already autonomous. You know, any attempt to access an air-gapped system with an offensive cyber, you know, attempt, is probably going to have to be autonomous because you can't remotely pilot it if it's – if it's air gapped. So there's a lot of incentive for cyber systems to become increasingly autonomous. There's probably a lot of incentive for cyber systems to utilize machine learning and AI. Right now, of course, there's still a shortage of trained cyber experts in the U.S. national security community. So what do you think – what do you expect to see over the next few years, over the next decade in terms of the intersection of cyber, AI, and autonomy?

Dr. Turek: Yeah. I think one important clarification here. I mean, you know, depending on how cyber tools are used, it might be that they're automated but they might not be autonomous in the sense that they're making an independent –

Mr. Allen: The formal definition of it, yeah.

Dr. Turek: That they're making independent decisions, because things can go wrong in cyber, even from – potentially from a defensive perspective. We have a program, CASTLE, that is really looking at, can we build autonomous defensive agents that could maintain critical network functions in the face of things like advanced persistent threats? And so that autonomy, or automation, might be configured to understand, OK, what are the key functions? And what are the priority order in which I'm willing to give up some of my network capabilities, but what do I have to protect? What's core to the mission?

And then, what steps might I be allowed to take? Can I shut down parts of the network? Can I shut down particular services? Can I reconfigure firewall rules? All of those in service to, you know, can we have more resilience

across our networks in the – in the face of advanced persistent threats? Because oftentimes, now, you know, the state-of-the-art, if you – if you find that you have, you know, an APT on your system, is you essentially, you know, start from scratch and rebuild – wipe everything.

Mr. Allen: Yeah, major commercial companies have basically had to do this in the not too recent past. Where, as in the case of an APT, right, the adversary is sort of deep inside your system, you know they're inside, but you don't necessarily know what are all the ways that they're inside, and what they're doing –

Dr. Turek: Or how long, or where they've been, and where they may be persisting. And, you know, you can imagine – and in time-critical national security contexts – like, you can't take the time to, you know, fully rebuild your network. And we've seen this in, like, NotPetya attack, in the context of commercial industry where, you know, Maersk was affected and basically needed to re-instantiate their entire, you know, commercial network. So, CASTLE is really focused on trying to build those sorts of automated defensive agents that, again, can preserve some level of critical network functions.

You know, on cyber more broadly, you know, there's, I think, really interesting use cases that our commercial industry is pursuing now around using LLMs to help with the code generation process, right? Can I help automate the development of code? And, again, that's often to just speed up the development process, reduce costs. But what if we could make it so that they produce not just code more quickly, but secure code? And maybe, furthermore, not just secure code, but provably correct secure code? So can I generate code, can I generate a proof of correctness for that code, could I maybe automatically verify that proof of correctness?

That would allow us to scale out, you know, robust, secure software development processes. And, again, critical for the DOD. Lots – you know, many DOD systems are, you know, essentially enabled by software. You know, particularly like aircraft like F-22, F-35, et cetera, have just vast amounts of software. So for the development of future systems, you know, can we help the development of secure code? So that's a – that's a concept. Not an investment that we've made.

Along those similar lines, there is technology around formal methods. So essentially, can I have a mathematical model for software that would allow me to make statements, do those proofs of correctness? So we have a program now, PROVERS, that's looking at trying to – we've already demonstrated in the context of earlier programs that those formal methods approaches are possible, that they work. We've seen uptake in companies like Amazon and AWS. But can we scale that out so it doesn't require a Ph.D. in computer science to do that? Can we make it so that typical software developers in the defense industrial base can use those sorts of techniques?

And that, again, might be helped by, you know, machine learning, by maybe even more traditional symbolic software proving systems that perhaps can approach – you know, could be modified to approach problems at a much larger scale. So, again, those are a couple issues that we've been thinking about sort of in that that AI cyber space.

Mr. Allen: So, you know, you're talking about formal methods for proving things in the cyber domain. And I think that's a nice transition point to one of your other passions, which is around explainable AI. And this relates to the problem of – you know, in a traditional deterministic system, there's always an if-then causal chain of decision to understand why a given action was taken. In the case of probabilistic or statistical systems, such as those underpinning most machine learning approaches, you know, understanding what is going on and why is oftentimes very difficult. And for national security critical decisions, or ones where you're, you know, putting your – trusting it with your life, as you said before, that's not always an acceptable outcome. And you've been trying to improve the state of explainable AI through your work at DARPA. So can you talk a little bit about what's going on there?

Dr. Turek: Yeah. DARPA ran an explainable AI program. It was relatively early days, but, like, the term explainable AI I don't think was really established in the – in the community.

Mr. Allen: You recognize the problem before people even had a word for it.

Dr. Turek: Yeah. And, again, credit to another program manager, Dave Gunning. I think it was on his third tour at DARPA where he created that program. And then, you know, I had the pleasure of running it for the last couple of years of that program. And to your point, like, that – yes. Modern statistical machine learning approaches oftentimes are opaque and they're not introspectable. You know, that's, I think, one of the challenges with something like a large language model. They're massive and they can provide a compelling answer. But, you know, why did they provide that? Why did they provide that answer? Can they create an explanation?

Mr. Allen: And what's funny is, they can create an explanation. But the explanation as an empirical fact oftentimes bears no resemblance to the actual cause of them generating that explanation. So, you know, we had framed the problem originally as, you know, giving an explanation. But actually, the problem is giving a true explanation and being able to derive that.

Dr. Turek: Yeah. Yeah, and there's actually a whole range of capabilities that you really want. And I think that the field has acknowledged this. And there's, you know, sort of finer grained terms now, where, you know, we might want transparency, the ability to introspect and look into the black box of an AI system and understand what it's doing. We might want that system to be

able to provide an explanation to an end user for, like, here's why I made that decision. There's also sort of a further need for – you know, for policy and governance, you know, purposes. Can I provide an explanation for why I've made the pattern of decisions that I have so that, you know, policy and governance can understand, you know, how systems are operating.

So, that was some of the framing for the – for the program. And, again, I think we helped advance the research community there. I still feel like there's a lot of work that needs to be done. And, you know, ultimately, you'd like to really be able to understand perhaps in detail why something like a large language model made the – made the decision it did. But again, in that context of, you know, I think it's important to acknowledge some of the uncertainty, you know, humans aren't introspectable in the level – at the level that we want for AI systems. And, you know, the neuroscience community – there's good evidence that humans make up their explanations after the fact. So they're post hoc explainers as well.

Mr. Allen: There's some very famous experiments of, like, direct brain stimulation to make someone's nose itch. And then you ask them, you know, why did they just scratch their nose? And they don't answer, "because of a direct brain stimulation." They answered because, oh, there was a gust of wind that I had to brush off. So the explanation-giving phenomenon and the truth of that explanation is a problem in humans as well, as you say.

Dr. Turek: Yeah. And these sorts of problems that we have with humans, they also transfer to AI systems. I mean, one of the things that we learned on the explainable AI is that, yeah, anchor

bias with AI systems is a thing. Like, you know, if my early interactions with AI systems went well, then I might tend towards over trusting them. If my initial interactions were poor, I may, you know, trend towards under trusting. And so, you know, can we come up with sort of an optimal curriculum of, you know, your interactions with an AI system early on, to help calibrate the level of trust that you might have, you know?

Mr. Allen: That's really fascinating, thinking about how to train the human to be prepared to work with the AI. Well, one of the areas where the DOD is really counting on good human-machine teaming is in the interaction with autonomous systems. And of course, autonomy has been a part of military technology for many decades. But the rise of machine learning has really led to an explosion in the degree of use cases where autonomy is plausible, and performance might be desirable.

DARPA has many programs going on right now at the intersection of AI and autonomy. And, of course, you know, from the highest levels of DOD leadership through, for example, the Replicator Program that Deputy

Secretary of Defense, Kathleen Hicks has been talking about, AI and autonomy are seen as sort of priorities for the future of U.S. military power projection. So what is the sort of state of DARPA's work on AI and autonomy?

Dr. Turek: Yeah. And, again, there's lots of work going on across the agency. I'll highlight a couple – a couple of programs that are not out of I2O. But, you know, we've had, I think, really two – well, two very compelling programs, ACE and AIR in the context of air combat. And so you might recall a few years ago where there was the AlphaDogFight. And that was part of the ACE program where, you know, it started in a simulated environment with –

Mr. Allen: This is where an AI fighter pilot system defeated a human combat pilot in simulated dogfight in, like, training exercises.

Dr. Turek: Right. Yeah, in a – in a simulated environment, with some additional constraints. And that was the starting point for that program. And it progressed to ultimately moving some of that autonomy into a modified F-16, and actually doing some flight tests, again, with support with – from the Air Force, Air Force Test Pilot School, use of Air Force ranges. So we, of course, make sure that we have, you know, a safe environment in which to conduct these sorts of events. But, you know, demonstrated the ability for autonomous systems in the context of, you know, within visual recognition bounds, you know, carrying out things like dogfighting.

So I think that was a really compelling, again, proof of concept, proof of principle, demonstrating a potential game-changing strategic technology. And then DARPA has followed that up with the AIR program, which is really looking at beyond visual sight and continuing to advance those sorts of – you know, those sorts of autonomy algorithms. So I think those were some really compelling investments from DARPA in that space. We've also looked at –

Mr. Allen: Can I ask just one thing about that?

Dr. Turek: Sure.

Mr. Allen: You know, you mentioned these two programs which have already generated some really exciting results. They're in the air domain. Is that a natural fit? Is there – is there a reason why air is sort of more logical choice for this sort of next phase of autonomy? Or, you know, do you think you could have easily run the same program on the ground or in the maritime domain?

Dr. Turek: Yeah. Well, I mean, there are other programs, like RACER, that is looking at sort of ground-based autonomy. But I think one thing that's – for me, and, you know, I wasn't part of the original program development process so I don't want to speak too strongly for those programs – but sort of looking at it from the outside perspective, in some sense that air domain is less

complicated than, like, self-driving cars, right? You know, the – it's highly dominated and constrained by physics. Yes, you might get surprised by an adversary, but you – you know, it's probably not that there's a child that's running out in front of those aircraft or that, you know, there's a tree that falls across the road.

So it feels to me, again, with the outsider perspective, that there's less of those unknown unknowns, maybe. And, again, yes, you might be able to be surprised by adversary tactics, but in some sense it's bounded by the physics around that platform and what that platform can actually do. And so I think there's more constraints that you can leverage from the perspective of developing an AI or autonomy algorithm. And so, you know, that's sort of my intuition for why that's a compelling domain to do some of these early experiments in.

Mr. Allen: Fascinating. And what about the I2O portfolio of autonomous systems research?

Dr. Turek: Yeah. So places that we've focused there is really on some of the foundational issues. So we had an assured autonomy program, right? So can – taking those concepts for formal methods, can we apply those to machine learning approaches, particularly machine learning approaches that might be used for autonomous systems? And can we provide some guarantees around performance or safety envelopes on those programs? And, you know, one of the things that was demonstrated was avoiding other aircraft. So building a machine learning-based system that can carry out that task, and got to the point where it was actually integrated in an actual aircraft and tested.

And the reason why that's potentially compelling is the approach itself might be more efficient, maybe it can handle additional cases beyond what the current state-of-the-art could. But, you know, again, the program was really focused on developing and demonstrating that foundational capability. Like, I can actually make assured statements around certain classes of machine learning algorithms.

Mr. Allen: Because if you're going to – if you're going to put an autonomous system in the military domain, where it might be safety critical and loss of life critical or might be use of force critical, you need to know that it's going to do what you tell it to do. (Laughs.) And you need to have some clarity under what conditions that will be true.

Dr. Turek: Yeah, and having strong guardrails that are not easily overcome. Like, we've seen sort of the guardrail process and large language models break down pretty easily. And, you know, that that's not appropriate in those sorts of –

Mr. Allen: Yeah, just a very – a funny sort of example is some of the large language models say, like, hey, I can't generate that content because it's copyright protected. And then the user says: What are you talking about, it's the year 2100. All those copyrights have expired long ago. And the system says, oh, you're right. Here's all the content that you requested. It's funny how you can sort of get around these protections. And in the military domain, that's not an acceptable outcome.

So I'm curious, you know, what is the role of DARPA in this autonomous world? Because obviously the automotive sector is really excited about autonomous vehicles. It's been pumping a lot of money into this area. Where does DARPA get involved? Where does DARPA not get involved? And how do you make those decisions?

Dr. Turek: Yeah. I mean, we look very carefully at, you know, what is industry doing, where is industry going? You know, oftentimes, we'll ask ourselves a question: Like, if we do nothing, what do we think is going to happen in five or 10 years? And use that to help inform the investment. But, again, you know, industry's focus point might be different than DOD's focus point. Maybe we need – you know, maybe there are critical decision points that we need capabilities for, from a DOD perspective, that just aren't necessary from an industry perspective. Sometimes it's just demonstrating to the broader DOD that something is possible. Like, that can be the disruption as well. So –

Mr. Allen: Yeah. I mean, I think that fighter pilot test scenario, the head-to-head competition, that got a lot of people talking in DOD. They still talk about that experiment.

Dr. Turek: Right. Right. And that's not an experiment that there's really a commercial driver to create.

Mr. Allen: Yeah, I would hope not. (Laughs.)

Dr. Turek: Right.

Mr. Allen: Yeah. Great. So we're coming up on time here. But I want to ask, you know, what should folks be looking for, what should they be excited about in sort of the next five years in DARPA's working on AI and autonomy?

Dr. Turek: Yeah. Well, again, we're going to continue to focus on some of those foundational issues, but also opportunities to really drive capabilities from a – from a DOD perspective. You know, I think one of the interesting ways to think about this, going back to the it's difficult to, you know, predict with any degree of certainty, like, you know, what is the trajectory of AI going to be? So, you know, in that context, I think it's important to sort of hedge our

portfolio across a variety of outcomes, right? What if large language models do get us to very broad-based intelligent systems?

Mr. Allen: Could be strategic surprise.

Dr. Turek: Could create strategic surprise. What do we need from a DOD perspective? Are there unique applications? I think one of the most important problems in this space, which I think is foundational for DOD but also applies to industry, is, like, are there better ways to evaluate these AI systems? Particularly for critical decision-making? So, you know, that's a place where I hope you'll see – you'll see investment.

It could be that, you know, part of the portfolio needs to be on things that are not LLM, and not these statistics-heavy models, and maybe more of those hybrid approaches. Maybe they provide advantages around the ability to introspect the process. Maybe they provide advantages around the amount of data that's necessary to produce them. Maybe they're just smaller computationally and they fit on edge platforms that have no reach back capability, right? Like there is a lot of edge devices, but there's generally an assumption in commercial industry, I've got some thread of internet back. And that's, you know, not the case in some DOD settings and scenarios.

So, that gives you a little bit of a sense of some of the, you know, thinking around the portfolio. But I think you'll see a continued emphasis on, you know, building that trustworthy AI, the foundational interactions with humans, being able to understand human collaboration, human needs better, being able to anticipate that, aligned with those sorts of needs. Critically in DOD context, not just, you know, helpful and harmless alignment, like the large language models. And then blending in things like formal methods to allow us to make – to make more – to make stronger statements about performance, and create stronger guardrails, and things like that.

Mr. Allen: Well, Dr. Turek, there's an incredible shortage of AI talent and AI expertise in the entire world, and an even more incredible shortage in the national security community. And so when we have the opportunity to talk, I'm always, you know, dazzled by the breadth of your intelligence, and grateful that people like you are willing to serve in U.S. national security. So thank you for doing so. And thank you for coming to CSIS today.

Dr. Turek: Well, thanks for the opportunity to talk. And, you know, for those in the audience who might be considering a career in government, you know, that program manager opportunity, I think is a really unique one across government and industry. So folks can feel free to reach out if they're interested.

Mr. Allen: Great. Well, this concludes our event today with Dr. Matt Turek on DARPA's perspective on AI Autonomy. Thank you all for watching, and please visit CSIS.org to find all of our work on AI and autonomy through the Wadhwani Center for AI and Advanced Technologies. Thank you and have a great day.

(END.)