Penn State Journal of Law & International Affairs

Volume 7
Issue 3 *Symposium Issue*

April 2020

Autonomous Systems &Emerging Technology

William Casebeer

Kevin Chan

Brian David Johnson

Patrick McDaniel

ISSN: 2168-7951

Penn State
Journal of Law & International Affairs

2020 SYMPOSIUM ISSUE

AUTONOMOUS SYSTEMS & EMERGING TECHNOLOGY

Moderator: Sara Rajtmajer
Panelists: William Casebeer, Kevin Chan, Brian David Johnson, and Patrick McDaniel

Sarah Rajtmajer: Good morning everyone. Welcome to our first panel on Emerging Technologies in Autonomous Systems. My name is Sarah Rajtmajer. I'm an assistant professor in the College of IST and faculty at our Rock Ethics Institute. My own work broadly is in privacy and security. I'm really interested in particular in understanding human social behavior and its impacts on security. It's a pleasure and honor to be part of this discussion this morning. We're very lucky to be joined by our four panelists who represent academic-industry, and government perspectives and whose day-to-day work covers the spectrum from hands-on design and development of new technologies to envisioning science-fictional future worlds. Allow me to introduce them briefly.

William Casebeer is director of the Beyond Conflict Innovation Lab where he leads development of science and technology to aid in conflict prevention and resolution. In particular, his lab leverages research in brain and behavioral sciences to inform peacebuilding and design interventions that

measurably promote social change. He is a retired lieutenant colonel in the US Air Force. He served as an associate professor in the US Air Force Academy, a fellow in human rights policy at Harvard's Kennedy School, a DARPA program manager and a senior research area manager in human systems and autonomy for Lockheed Martin's advanced technology laboratories.

Sara Rajtmajer: Next to him, we have Brian David Johnson. Brian is a professor of practice at Arizona State University School for the future of innovation in society. At Arizona State, he heads the threat casting lab whose mission is to envision possible threats 10 years into the future. He is also a futurist and fellow at Frost and Sullivan, a consulting firm focused on innovation opportunities driven by disruptive technologies. In that role, he consults with governments, militaries, trade organizations, and startups to help them create strategies that embrace emerging disruptive technologies.

Next to him, we have Kevin Chan, an electrical engineer and network science team lead with the computational and information sciences directorate at the Army Research Laboratory (ARL). Within ARL, he works on specific programs in network science and cybersecurity. He holds degrees in electrical and computer engineering and public policy. He is published on game theory, network science, complexity as well as privacy, cryptography, telecommunications, military computing, and command control.

Finally, we have our own Patrick McDaniel, a distinguished professor of computer science

and engineering and the William L. Weiss professor of information and communications technology in the school of electrical engineering and computer science here at Penn State. He directs the Center for trustworthy machine learning which is a frontier project funded by the NSF and consisting of faculty from across the country.

Sara Rajtmajer: The goal of that center is to develop safe machine learning robust to attack that can provide a basis for the application of intelligent algorithms in new domains. Dr. McDaniel has served as a program manager and lead scientist for the Army Research Lab Cyber Security Collaborative Research Alliance and prior to joining Penn State with senior research staff at AT&T Labs. Our distinguished panel is here to talk to you today about the future of autonomous systems and emerging technology. We have the privilege of opening the day's events, so I thought I would start by saying just a couple words about where we are with AI and autonomy and perhaps how we've arrived here.

There are more people talking about AI today than ever before because the first two decades of the 20th century have brought us striking examples of what's commonly referred to as autonomous technology and artificial intelligence, self-driving cars and drones, robots in deep sea and space exploration, weapon system, software agents such as bots in financial trade and deep learning and medical diagnosis are just a few prominent examples. Given that recent attention and the hype and speculation, one might be forgiven for getting the impression that AI is this awesome new

idea that's just emerging, but it's also important to think about the history and the context for our discussions today.

Sara Rajtmajer: As the field of scientific inquiry, AI traces back to a conference at Dartmouth in 1956 where John McCarthy brought the term artificial intelligence into the vocabulary and prior to that, many of you are familiar with Alan Turing's intelligent machines. Perhaps a useful way to think about the history is DARPA's three waves. DARPA has divided the history of AI since 1960 into three phases. The first wave handcrafted knowledge, so experts took their domain knowledge and characterize it in rules that could be fed to a computer, and that computer could study the implications of those rules.

Those are examples scheduling systems, even your tax software, but also is relevant today and in DARPA's grand cyber challenge, the winner of that challenge actually used what we would classify as first wave knowledge. It's certainly not outdated, but first wave AI suffers in the real world. Many domains have moved to second wave AI where engineers create statistical models for specific problem domains and train them on big data. This is most of what you think of today as AI. This type of second wave AI is behind voice recognition and face recognition, and second wave technologies have been awesomely successful in classification and prediction given sufficient data, but with learning, skewed training data can cause mal-adaptation and generally, second wave systems lack contextual and reasoning capabilities.

This is where the third and final wave comes in, and I think really much of what we're discussing today in terms of looking forward into the future is really in this third wave, which is a vision really more than a reality of a future AI where these autonomous systems can construct contextual models for classes of real-world phenomena, and that context can inform the ability of the system to reason and to explain.

Sarah Rajtmajer: What we'll try to do here in the next 70 minutes or so is focus on this third wave, what it looks like and in so doing, I hope that we provide ground for the rest of the day's panels and events so that is we'll look into the crystal ball a little bit, lay out what might be coming in AI and autonomy in the next five to twenty-five years and discuss ways in which emerging technologies will impact security. I've asked the panelists to start by just preparing a few remarks on what they see as key trends in AI and autonomy in the next five to twenty-five years so a broad task, but we'll get everyone's perspective to start and then from there, I've prepared a few questions for discussion, but I'll also leave time for everyone to have the opportunity to ask questions to our panelists at the end.

Anyway, with that, let me turn it over to Bill Casebeer to begin.

William Casebeer: Okay. Thanks everyone for generously donating some of your time this morning, to Sarah for the great introduction and to Admiral Houck for convening the conference and for the invitation to speak. In the Beyond Conflict Innovation Lab, we've been trying to apply AI

and autonomy in general to help develop technologies that generate positive social change and aid in preventing conflict or deescalating that when it happens, and really the only reason we can even think about something like kind of that mission is because of these developments in the third wave AI that Sarah set up with her introductory remarks. What I want to do is make three general points and then turn it over to Kevin.

William Casebeer: My first point is that one development we should be tracking closely in autonomy is the role that human state assessment and sensorization plays, and enable ones to build effective human machine teams. There's been lots of developments in that technology in the last two decades especially and as our keynote speaker, Paul Scharr, mentioned last night centaurs, human machine teams are I think really the future of autonomy and that's because in general, warriors work in teams. As we think about national security and autonomy, I think we need to give serious thought to how it is that we create effective human machine teams and how we endow the autonomous pieces of that war fighting system if you will with the eyes, ears, and brains they need to understand what their teammates are up to, so they can adapt accordingly.

In that regard, I think the most critical development in the last twenty years has been the launch of a whole host of human state sensing devices that go even beyond what you and I can sense with our eyes and ears, so that includes things like maybe devices you're wearing right now. How many of you are wearing Fitbits or something somewhere or an

Apple Watch? Okay, so looks like about 30%, 40% of the crowd, right? That's a human state sensing and monitoring device and there's some algorithms behind it that let it interpret maybe your heart rate or potentially even your heart rate variability data. It was that second stream of data that Apple has used to develop a predictive algorithm to tell when you might be having a heart attack, for instance, that you may have seen in the press.

William Casebeer: That's just the start of the sensing you can do to help adapt an autonomous system to the state of its human teammate. Another one that you could potentially use is electroencephalograms and so that's just something that measures brain waves coursing over your scalp every moment of the day. Here's an example of an EEG device. This is a commercial device that professional meditators use to assess their brain states. They look for suppression in one frequency band of those electrical patterns on top of your head called alpha and if you can suppress alpha wave, that's indicative of your ability to enter a transcendental or meditative state.

They use a biofeedback paradigm to train you implicitly to push down your alpha wave. I can actually just put this on and show you what real-time neural state assessment looks like. I had two reference electrodes in my ears, four passive electrodes that sit on my scalp and then I will try and relax my face, and I'll hold up the iPad display here. I'm going to have to stop talking for a moment, but as this electrode settle in, you'll see some squiggly lines. Now I have to be still. Those are real-time readouts of

some of those frequency bands that I was talking about earlier.

William Casebeer: Alright, so this is a $200 piece of technology that you can use to monitor human state and that data can be fed back autonomously to help adapt the human machine team and in my previous lab at Lockheed Martin, that was some of the work we did. We used EGG and a host of other sensing methodologies to do things like workload assessment, so can I have an autonomous algorithm that redistributes tasks amongst a human team to improve the performance of the team? That's the first development I think we should track, human state assessment and how that impacts centaurs human machine teams in the development of autonomy. Second quick note is I think we do have some concerns as we do that.

The principal ones that keep me up at night are ones of transparency and intelligibility. The nice thing about human to human teams is if I have a question about what you're doing, I can always just ask and, more often than not, if you were reasonably well put together human being, you can at least offer some insight into why you took the action you just did. "Hey Brian, why are you speaking so loudly?" "Well, it's because the microphones are failing." "Oh great, maybe I should speak more loudly too," right? That's a typical interaction you'd see on an effective human-human team.

For human machine teams, I think we ultimately need that same kind of transparency and intelligibility because if we don't have that, not only will the team be ineffective, but it will also leave us open to exploitation, especially of

some of the representations, heuristics, and biases we use to reason about the world as we tackle it. If any of you have been following that literature, it's very interesting. There was a fashion designer three years ago, for instance, who developed a set of scarves that used some eigenvalues, technical term forgive me, that were extracted from the interior layers of a machine learning network that had been trained to recognize faces and she smeared them across a scarf she could put her on her neck.

William Casebeer: When you point that face recognition algorithm at the scarf, it misidentified her as having hundreds of faces around her neck, even though it just looks like a series of dots on the scarf, so kind of that failure to understand how these systems how they represent and reason about the world can leave us open to 21st century forms of cognitive camouflage, concealment, and deception that I think will be an entirely new feature of the security landscape in the 21st century and beyond. You probably saw if you had your Google feed queued up yesterday, the team that developed stickers that can be placed on the road to cause autonomous vehicles to think they are steering in the center of the lane when they actually aren't.

Paul mentioned last night, that uneasy feeling you have when you felt the car jerking it back to the center of the road. The designers of those systems, I can't say for certain may or may not have a notion of what internal representations are being used by the car to let it judge where it is in the world and if those kinds of representations are laid down in the

road, they may look like nothing to us except a series of dots or a strange looking square, and yet our autonomous vehicle might interpret those as being "oh, the road is turning to the right." 21st century cognitive camouflage concealment deception is critically important.

William Casebeer: My final remark is about the opportunity that autonomy developments create for us and controversially, provocatively, hopefully I think one great opportunity they give us is the ability to develop a truly artificial conscience. When Paul talked last night, his next to last concern was about the morality autonomous systems and I think we need to be open to the possibility of developing systems that can reason in the moral domain as effectively or even more effectively than their human teammates. We definitely have a need for this because as Paul mentioned last night, we have these systems even presently that are making decisions that involve the release of weapons and the intentional harm to hopefully non-innocent people to combatants in the context of war.

If we're to be an effective human machine team with our autonomy in the future and not put ourselves at a competitive disadvantage to militaries that develop these human machine teaming systems, I think we do need a moral governor for our systems. The approach we could use to build an artificial conscience isn't that different from the approach we use to build a natural conscience presently, right? If you have children, you're raising them and you're working with their very plastic brains to help train up a neural network architecture that ideally is going to embody the ability by the

time they are full-blown agents themselves to make good decisions about what constitutes a flourishing life, about what constitutes good habits and dispositions, and about what actions tend toward good consequences.

William Casebeer: You're already doing this with your children and with your friends and peers. Let's work on the formalisms that let our autonomous teammates do that, and there could be any number of the machine learning techniques, deontic logics, traditional first-order predicate calculi. There are lots of people who are working in the domain of thinking about how we could build an artificial moral reasoner, and the content from that could come from the three grand traditional moral theories that drive a lot of our moral actions implicitly. I call those the three Cs, considerations of character, consent, and consequences. Two of them Paul mentioned last night in his presentation.

The idea here is that if your artificial conscience reasons about what kind of habits and dispositions they'd ought to develop to be a good teammate, that's character development for the system, if it reasons about human rights, what actions it is absolutely prohibited from taking, what *mens rea* if you will, or state of mind, it ought to have. It's thinking about deontic or duty-based concerns and then finally, if your AI's conscience is thinking about future consequences and what actions it can take to produce good ones rather than bad ones, good old John Stuart Mill in action, then it's reasoning about consequences.

The integration of those three things I think presents lots of opportunity for us to develop

yet more effective teammates who can help us make war when it is necessary as morally permissible as it can be given its nature, and I'll stop there. Thanks very much.

Sarah Rajtmajer: Thank you. Thank you.

Brian David Johnson: Do you want me to go?

Sarah Rajtmajer: Sure.

Brian David Johnson: All right, okay. Well, good morning everybody. It is a pleasure to be here and I'd like to if you indulge me start with a quick personal reflection that I think we can apply to all of the great panels that we're going to have today, and then I'll get into real quick the work that I'm doing. This is my first time at Penn State. I'm super happy to be here. It's a beautiful campus, I absolutely love it. Also I'm so happy to be here because my family is from Pennsylvania. My mom is actually from a little town that you've never heard of called Gouldsboro, Pennsylvania, which is about two and a half hours northeast of here smacked dab in the middle of nowhere, and I spent a lot of time up there when I was a kid.

I was here and I was walking around the campus just reflecting that Penn State, a beautiful school that my mother went to study computer science engineer, but she didn't because she was a female in the middle of the 20th century and she was poor. She didn't get to come here and it took her 20 years to get her engineering degree. Don't worry, her story is very common, but it's a tragedy. Now it took her 20 years and ended great. She ended up getting multiple engineering degrees. My mom's a great engineer she had a great long

career in the US government, so it was great. As I was walking around the campus yesterday and actually just imagining that I started thinking about the laws that the practice had from an entire generation of minds of young brilliant engineers that we lost.

Brian David Johnson: What could they have done to the PC revolution, right? We needed their enthusiasm, we needed their passion, and we didn't get it. I was thinking as we think forward in myself being a futurist thinking of as we start to tackle these very, very hard problems that are coming, and they are coming, and they are complicated, and they are big, and they are going to affect every single human being on this planet, that we need as many people working on these as possible, and we need a diversity of gender. We need a diversity of background of domain. I love actually what people are saying about making sure we're getting as many different domains, but we need to be actively inclusive because to solve these problems, we know that homogeneity makes brittle technologies.

It's only through being diverse that we can create robust solutions, not just robust technologies, but robust solutions. It's one of the things I think as we think about the future and where things are going to know that there's always going to be people who aren't included. Now certainly luckily last night, I was able to sit and chat with student Megan. I can't see her if she's here now, but I was talking to Megan last night. We have brilliant young ladies who were here now and it made me text my mom last night, I was all proud of her. It was very late, so she's like, "Why are you texting me?" That has changed and we have made progress,

but we can make so much more progress because we need that diversity of background, that diversity of domain, of ethnicity, of gender.

It's very, very important so that's one of the things that I push is we think about the future of autonomy saying are we constantly, not only as leaders, many of you have gray in your beard like myself, so we've been doing this for a while, but even to the folks who are just starting their careers, are we being actively inclusive, are we creating the requirement that we're actually getting as much big and a diverse team as we can because we can use that enthusiasm and that passion and that diversity to actually make not only better technologies and better solutions where we can do it to make a better future. Thank you, so thank you for indulging me on that. Hey everybody, I'm a futurist.

What I do is twofold, so I work with corporations, so I work on the private side of a private practice where work with basically Silicon Valley as well as manufacturing ag, do a lot working FinTech as well as medicine. Over the last five years, I've seen that shift beginning to happen where we're starting to see that shift where people are starting to make industrial grade artificial intelligence, and so that's one of the ways that when I talk about artificial intelligence and autonomy. For me, I call it autonomy whether it be digital autonomy or physical autonomy, it's autonomy. I've been seeing people actually making industrial-grade autonomy and when I say industrial-grade, what I mean is not smarter than human AI, right?

That's just very small sliver, but this industrial-grade that just does work, right? It helps you land the plane, it helps you pick a movie, it helps you do all this work, and I'm seeing in industry, this get applied more and more and more and more. What's interesting and I think in the perspective that I bring partially to this panel is to say so working in house with these folks. I was the in-house chief futurist for the Intel Corporation for over a decade and now I work with a lot of other organizations to do this, is we look 10 years out and what I see in this 10- year time span is we're beginning to see more and more industrial-grade AI put towards business use, which I think is very different, very different from I think some of the conversations we might have about national defense and national security that in the corporate realm, it's all about shareholder return and it's all about how to get tasks done.

Brian David Johnson: Only recently, and I think we're going to be only seeing more and more use of that, but it's more and more specific use around business ROI and business rules, and we're going to see a lot more autonomous technologies working behind the scenes to actually go and create better experiences, create better one-on-one, and you'll see even more personalized AI coming in. One of the things that I've seen in the last couple of years and I think we all have seen is some of the hazards that have cropped up in the private sector where you can no longer say which hopefully you've all famously seen these people say just get up in front of congress and say we never thought people would use our platform to do that.

This is a point from which you can never come back from. Now you're beginning to see industry start to make that shift and start looking at these ethical concerns and start looking at the application of this. I think also in the next five years or so, you're going to see more and more of that whereas we have the majority of the autonomous work being done in the private sector and we're starting to see the private sector start to catch up with the public sector in that area. That's one area over the next five or ten years. One of the other things that I do is I run a threat casting lab at Arizona State University.

What we do there is we look ten years out and model possible threats to national security, and then we turn around and look backwards and say, "Okay, how do we disrupt, mitigate, and recover from those?" Some of the findings that I think could be helpful for this discussion is one of the reports that we did was called the widening attack plane and we were actually talking about this at breakfast this morning. One of the things that we're starting to see and we're only going to see more and more when it comes to not only national defense but also when it comes to criminal actions as well, where we're not just seeing leaks or cyber-attacks, but we're starting to see cyber social attacks, we're starting to see cyber physical attacks, and certainly starting to see as we move forward, cyber kinetic attempts.

I know we were just speaking before that sometimes it's called hybrid warfare. We called it just blended attacks, that you're going to see these blended attacks and what is essentially the widening of the attack plane, so that's one

of the ways that we think about it in the threat lab is to say over the next 10 years, you will see a constellation of technologies widening that attack plane, that what these autonomous technologies will allow us to do is start to tap into the internet of things in smart cities and robotics and certainly a physical and digital autonomy, but it's not just one. It's all of those, that that actually becomes an attack plane and every single device becomes an attack surface, so we begin to see that widen and widen.

What that means is that no one actor can perform all tasks, so the government can only do so much, the military can only do so much, private industry needs to step up, academia may need step up. I think doing events like this is extremely important, so we're starting to see and I think this is a whole of society problem and certainly a whole of security problem that we need to actually work with people that the technologies themselves aren't that hard. It's actually getting everybody to work together which is actually the hard part and as a part of that, what it could riff off what he was saying before in both the public and private sector, I'm going to split hairs here when we talk about ethics and autonomy.

Brian David Johnson: I'm not a philosopher and I'm not a specialist in ethics, I'm an engineer and a futurist. When we talk about ethical AI or ethical autonomy, I actually think that we're having the wrong conversation, that for me, it's not about an ethical AI or ethical autonomy. It's about ethically compliant artificial intelligence because ultimately, we have to remember, it's about humans. I think this is the thing we forget and especially as we get into these

autonomous systems, that we're imbuing them with way too much, that ultimately, we need to understand that these are tools and these are tools that we create. What I do in the private sector as well as the public sector is I turn the light back and say I can give you the required document to show you how to make an ethically compliant artificial intelligence.

Again, I'm a systems architect, I can show you how to do that. You're actually just adding a couple of pages and a little bit of validation on the back end that we learned from human-computer interaction back in the 90s. It's actually quite simple. The hard part, and this was at the keynote, is what do we value. These are the conversations that we're starting to have and over the next five to 10 years, I think we need to have more and more and more of to say what do we value, why do we value it, and to actually have that plan. Then the final bit I'll leave you with as Sarah mentioned in the introduction, I'm also a science fiction author.

Don't underestimate the power of science fiction to scare the heck out of people, which actually this is what sci-fi does a lot and I like to admonish my sci-fi authors about that, but we have to make sure that we're telling ourselves the right stories about the future. That's one of the things that I've learned as a futurist over the last 25 years of doing this, is that the way that you change the future is you get people to change the story they tell themselves about the future that they will live it because if you can do that, they'll make different decisions. They'll make different policy decisions, they'll make different education decisions and business decisions that

those stories that we tell ourselves are incredibly important.

For each of you who are doing work in this whether you are engineers or you do work in policy or law or anything, especially when it comes to these autonomous technologies, what's the story you're telling yourself about this future. You have to articulate it, both the story you want and the future you want to avoid, but then also what are you telling, to your colleagues, to your students, to your children, to your parents. Those stories really, really matter and I think we need to be really cognizant of how we tell those stories. Thank you.

Kevin Chan: Okay, and as I was kindly introduced, I'm from the Army Research Laboratory and so we're tasked with looking at the 2040, 2050 timeline, so this is looking at basic research. In terms of looking into operational or military types of questions, I think I'll defer that to some of the leaders or the organizers of this event who have had distinguished military careers, but in terms of technology development and capability development, yes obviously AI machine learning has become a focus of Army and DoD interests.

I'm from the Computational and Information Sciences Directorate at ARL and there's a whole host of folks that are doing a lot of robotics and autonomous systems research and a lot of human-computer interaction research. I'm coming from more of a network science and cybersecurity perspective. Some of the work that we do is collaboration with Patrick who's next to me, so I'll defer some of that

work for him to explain. In terms of this panel, the question is what will future warfare look like or how do we envision it, and I'll reference one document that was recently published by TRADOC and this is what you're talking about, this blended type of operations.

Kevin Chan: The document is called the multi-domain operations, and this is looking at the different domains of warfare that the previous operating concept was air, land battle. You had to coordinate the air forces and then ground, essentially, you're not bombing the places where folks are and your own guys are. Now the multiple domains are sea, air, land which are the traditional domains and then you have space and cyber. Now the question here is the attack plane, but there's also the coordination plane or the control plane, how do you coordinate all these different domains and carry out missions. Oh, you're right, so there's the technical aspects of command and controlling all the elements that are involved.

You have a lot of military personnel involved, you have a lot of materiel involved, and then obviously there's policies and strong opinions by your military leaders that want to have command of their information, their assets, and how do you do this and how do you collaborate and do something meaningful. One of the thoughts is obviously how does AI fit in this, right? I think a lot of the discussion here has been in terms of autonomous weapon systems, but if you're looking at AI in a broader sense in military operations, this will not just occur at the terminal points of operations.

This will have AI in the headquarters, will have AI bringing together intelligence from these different domains, and so maybe one question to ask is, if you have an analyst that's an autonomous system, would you take an order from an autonomous system? I would probably ask the commander or the folks with a command experience and they would probably say absolutely not. The question here is this man-on-man teaming is as Billy talked about, right? How do we work alongside and one another and leverage what good things that each of us can provide in operations?

Kevin Chan: Right, so I guess I would say that we need to look at autonomy more than just a narrow AI that can do specific tasks, but future thinking is can we develop general AI or autonomous systems that can do a broad range of tasks and will we adopt those or let those into our military organizations. I think this was also mentioned yesterday in terms of the Stuxnet and cybersecurity, but the idea here is that a lot of the decisions that we need to make are at millisecond speed and these are things that humans cannot do. Will we rely on autonomous systems to make these very quick but very important decisions and can it do the risk analysis and can we delegate that decision authority to these systems?

Then maybe a couple anecdotes that I'll mention to give the current state of autonomous systems. One that you may have seen, I don't know—I guess it's the computer science folks—there's an activity called RoboCup and it's basically robots playing soccer. Don't laugh, but they have different divisions. One just look like tiny little Roombas

that actually can play soccer very well with ping pong balls, but there's a division called humanoids and the goal of that event is to basically beat a human team within twenty years or something like that. I will tell you that you can look it up on YouTube. They have a very hard time standing up.

Brian David Johnson: You can't watch the edited version. You have to watch the whole version because it basically looks like this. No, I'm glad you brought that up. I make my students watch the unedited version for like an hour and they're like, "Please sir, make this stop." I'm like, "Yeah, now let's talk about the reality."

Kevin Chan: Another more operational anecdote is I was at a field exercise and I think the keynote mentioned it yesterday, what do you do when you lose contact with the autonomous system, do you let it shoot, but I suppose a more basic question is if you lose contact with it and it's not doing what you want it to do, how do you just land the asset? I guess I'll end this since we're near Beaver Stadium, so basically, we're at a field exercise and the default pattern for this UAV was just a circle around the field, and they're trying to figure out how to get the UAV down. Essentially, they've wheeled up the Allstate good hands extra point netting and had the thing fly into it—it was good.

Brian David Johnson: It was good.

Patrick McDaniel: Alright. Well, I guess I'm going to start with a brief story. Well, the larger event, Google every year brings the top 10 or 15 lab heads from across the country. We go to Palo Alto Mountain View and we spent about two days with Google, and they roll out the next

generation their skunk work of projects and about six years ago, I was sitting in the room and there's a bunch of us sitting there, and the Google brain team comes in and they start rolling out image recognition which is at the time would enter the realm of science fiction. It was really the point of inflection where the Google brain people really got good at image recognition and there was a colleague of mine from the University Wisconsin leaned over to me and he said, "This looks like magic to me."

Patrick McDaniel: I think the next five years of this technology has been one magic story after another and I think it's really important from a basic science standpoint to understand where we are with respect to the technology and what its limitations are and in a very fundamental way. I think the bottom line is we need to understand it's not magic. Machine learning and AI is not going to be more moral than human beings. Machine learning and AI, the reality is it always looks more sophisticated and more intelligent than it really is because what we're really good at in technology is simulating things that look like intelligence.

We don't really have anything that approaches intelligence and the reason for this is that a number of different approaches for reasoning and machine learning that was brought up this morning, all of them are really doing what we refer to as either reasoning under fixed notions or what we call generalization, and all of that really means is we're learning from examples that we can see or we're learning from or we're reasoning from axioms or statements that we make. The limitations of machine learning and AI is really the limitations of human beings. We

are not going to solve problems that humans couldn't do given enough time, and so there's a lot of consequences for this.

Patrick McDaniel: AI and machine learning is only good at things that seen before or have been anticipated. We heard yesterday during the keynote a number of examples where you have something that works in isolation, it works great in a lab and you put it in the field and it doesn't work. The reason it doesn't work in the field is what we have is called a domain shift. The classic example was there was a DARPA Grand Challenge for autonomous vehicles that goes back into the middle early 2000s and they had these autonomous cars. I believe it was in Pittsburgh and they're driving around a course and everybody's doing fine, and then a cloud comes over and one car just slams into a wall.

It's because the machine learning had never seen a cloudy day before. They had always done all of the training under sunny circumstances. AI doesn't generalize the way we do and so it didn't have any way of dealing with that or dealing with that domains shift. What AI and machine learning is really good at is things that are finite and controlled, limited tasks we refer to them in the science community, things like object recognition. We're talking about recognizing boats in a water where you have a missile going that's patrolling an area looking for a ship to hit. That is actually a fairly simple task with respect to the real domain.

You have an image and you have some algorithms for figuring out the edges of objects in that image and then you just figure out well

that's ship like, and so that is what AI and machine learning is good at. If you take a new ship that's perhaps round, it's a raft and the AI and machine learning hasn't seen it before, it's not going to recognize it as a ship, but we as human beings will immediately know that because we do what refer to as contextual thinking, and that came up a little bit earlier today, but contextual thinking is really hard because there are lots of different environments that you simply can't anticipate. Machine learning and AI is not really good at that now and I'm skeptical that we're going to get good at it in the short term.

Patrick McDaniel: The other thing is that AI and machine learning is absolutely terrible about ambiguity. One of the realities of things like morality and making tough decisions. We heard that story about the 6-year-old girl who was performing reconnaissance for enemy combatants, that is an example of ambiguity. That's a morally ambiguous situation where you are perhaps putting your own men in harm's way because you're not going to do something about the 6-year- old girl. That is not something AI machine learning is going to solve for us. There's a great example that I was talking to some folks at Harvard, so obviously Harvard is having some complex discussions about admissions right now.

One of the discussions we have with Harvard is should we apply machine learning for Harvard admissions, right? That's a great example where all of a sudden if we just have a machine learning algorithm to figure out what makes a good Harvard entrance, then we're not going to get sued because hey, it's the

algorithm who made the decision, but in reality, you can't just offload responsibility for tough decisions on AI and machine learning because it will only learn what you tell it. It will only learn from the examples and this leads into some economic theory. Once you start replacing important phenomenon like admissions into Harvard or deciding who gets a loan, there are all kinds of secondary problems.

Patrick McDaniel: We got into fairness a little bit and if I took home loans, the home loans that were accepted in Cleveland from 1970 to 1985, you would find that the African American community was substantially prevented from getting home loans in Cleveland. Now if I created a model using that data and I use that to decide home loans, it would just reproduce that systemic injustice into that model. Models learn exactly what you tell them, but there's a broader economic theory and it actually goes to Goodhart's law. Goodhart was a 1970s British economist and he came up with the Goodhart's law, which says at any time you create a metric, it's immediately bad.

A good example is miles per gallon, so miles per gallon is the proxy for environmental impact, right? That was created I believe in the late 1960s, early 1970s, so miles per gallon all the sudden, you can go to any car and there's a number and that number is the environmental impact of that car. Well, that's not really true because environmental impact is so much more than just how much gas it actually burns, but it became a proxy for environmental impact and in so doing, in creating this metric to try to make cars have less environmental

impact, it ignored all of the other factors like tire wear and road wear and weight and all the other things that go into it.

Patrick McDaniel: Now I would say that the same thing is going to happen as we introduce machine learning. As soon as we create an algorithm that says Penn State or Harvard is going to do their admissions by machine learning, someone's going to figure out that the model really likes people who play varsity task and all of a sudden, there will be tennis clubs growing up all over the state of Pennsylvania because if they take that training data, there'll be that inherent bias. People will figure out what makes the model happy, not what makes it good Harvard or a Penn State applicant. This gets to another point I think Bill made really well here is that also AI and machine learning is a consequence of that manipulation of the models.

What we need to know is that AI and machine learning is inherently deceivable. It does not reason the same way we do. It just implements the model that we produce for it. Great examples, historical cardboard tanks are great for vision systems, right? You put cardboard tanks, you can get people to waste missiles. We heard about adversarial patches. People can put patches on signs and makes autonomous vehicles misclassify, and I mentioned things like admissions. I think the broader question is we want to avoid the virtual cognition trap and this is really repeated constantly in the press, is that they imply through a lot of these articles, every time there's a new system, that there's some real cognition that we would understand

his cognition going on underneath the hood. There is not.

Patrick McDaniel: We are at least a century away from real cognition, and we talked a little bit about moral systems. When we're talking about moral systems, we're not talking about morality in our sense. We are going to be talking as Brian said about simulated morality. You put it as morally compliant, right? That's the simulation. If you give it rules, it'll follow those rules. You are not going to get the same kind of moral weighting that we do that we probably couldn't even articulate in a meaningful way and to the broader point for people in the law and policy space here, I think I would agree wholeheartedly with Brian is we're having the wrong discussion about policy when it comes to autonomy, not just in the military setting, but in the broader setting.

The thing to know is that AI and machine learning is what we refer to as probabilistic reasoning, right? It's making decisions based on the best information has and as a consequence, it will be wrong sometimes in the same way that humans are wrong sometimes, and there's this propensity for people to think that AI and machine learning is going to get you to 100% accurate system. It is a technical impossibility for a military system to recognize an object in a battlefield with 100% accuracy. We will never get there, period. It's just because the world, because the light physics are too complicated, because objects change, because environment's changing, you will never get to 100% accuracy.

Patrick McDaniel: Rather than have the current discussion which is well if it fails, can I deal with the consequences, the real discussion should be it will fail and it will fail pretty regularly if it's used a lot. We've seen this in autonomous vehicles. Most of failures in autonomous vehicles, you don't actually see because the systems are self-correcting. You don't have the major fails, but they're failing all the time. The real question we should really have is when they fail, how do we deal it, can we accept the consequences of that failure, and that's why things like the Pegasus systems are wonderful. I mean not Pegasus, centaur systems are . . . I was in the Greek somewhere.

The centaur systems are probably, I would agree with Bill, that they're probably the best-case scenario because we can retain the consequences of actions. Let me say just in the last 60 seconds or so. I think it's not hyperbole to say that we're on the cusp of one of the great transitions in our existence as a species, and I'll give you an example. There's about 8.2 million people involved in the trucking industry in the United States and we are entering an age where the trucks are going to become autonomous, the economics are overwhelming, right? If you know anything about business, you spend any time with finance people, people are super expensive, everything else is free, equipment, everything else is free, so we're going to go to autonomous trust.

Eight point two million people are going to be affected by the fact that most of the truck drivers are going to disappear, but that's going to have secondary effects on our economy. Think about flyover states in say Kansas where

the truck stops are one of the major employers in these small rural communities. There will be no need for a Stuckey's in the middle of Kansas any longer because there won't be any truck drivers. There won't need to be all of the other services. Gas stations will disappear because the trucks will pull in to fully automated refueling stations and the trucking industry will learn that it can be much cheaper rather than paying say Shell for the fuel.

What you're going to have in just trucking alone, you're going to have 8.2 million people or some large percentage of 8.2 million people being essentially pushed out of work, but the secondary effects, the cascading effects of no trucks on the roads is going to have enormous impact particularly on the already hurting rural America. This is just one example. You can look at finance industry. You can look at the insurance industry which really people move paper. You look at the education, people like Penn State's in 10 to 15 years is going to need a lot less people.

Patrick McDaniel: There's going to be an enormous social disruption to this technology, and so I think it's really important for us to understand that it's extremely limited and that when we deploy these systems, they're going to have negative consequences to the society at large. Thank you.

Sarah Rajtmajer: Okay, thank you. Okay, so it is 9:00, but because of what Patrick has just mentioned, I wanted to ask the rest of the panel about your thoughts on artificial general intelligence, right? It sounds like Patrick is saying one century, but there's so much hype where people in many

ways feel like this is around the corner. As Patrick mentioned, what we are seeing the second wave as I set up in my introductory remarks is something that looks like intelligence and we have two ways to get there. One way is learning from examples that we see and another way is by giving computers rules that we design, handcrafted knowledge.

Sarah Rajtmajer: Can I ask then the rest of the panel when we have these discussions about true intelligence or let's call it generalized intelligence, also maybe the third wave that I've described is another language for that, what do you feel like is the timeline and perhaps even if you could speak to the history as context for your reasoning? There have been these summers and winters proverbially in AI since 1960 and what are we in right now and what are we looking at, maybe is it a century or how do you feel like that could progress for us? Let's start with Bill.

William Casebeer: Okay, thanks Sarah. Yeah, rich set of comments. Thanks everybody, really fantastic remarks. I actually think that we do have some instances of domain constrained artificial intelligence that are implementing real cognition right now, right? If we think about cognition as being computation across representations which is the standard theory of what cognition consists in and if we think about useful cognition as being computations across representations that let us get things done in a given domain, then we're already there. Everybody on this panel myself included maybe has put in a credit application to buy a house or to get a car or something like that in the last decade.

William Casebeer: Chances are that initial cut against the credit application was never touched by human being, so there was I think bona fide cognition going on their judgments, reasons, representations with computations happening between them that help that system arrive at an initial judgment about whether or not I was credit worthy. I think that in some domains we do have real cognition taking place and that we're already subject to it as part of a human machine system even now. I would even go so far as to say there are existence proofs of moral reasoning systems taking place right now as well.

I actually built one as part of my dissertation work where I trained neural network to take Lawrence Kohlberg's defining issues test and was able to get it to pass some of the standards that we hold ten-year-olds to in a very limited domain. Totally, lots of brackets are on this claim but that nonetheless exhibits some of the same dimensions of reasoning that we do when we reason about moral issues. I think in some respects were already there. I think newfangled cognitive approaches like neural networks and connectionism can help us build general-purpose reasoning systems. We have tenth to the fourteenth neurons at least with tenth to the fifteenth connections between them and our three-pound universe on top of our spine.

By the time we're eighteen, we've been exposed to millions of hours of training against millions of exemplars and thousands of tasks and context domains. I think if we give our artificial systems enough time that they'll get there too. I don't know whether it'll be in the next few years or century from now, but I'm confident

that with the kind of work that's taking place in labs like Matthias Scheutz's at Tufts, where he focuses on building theory of mind systems to let us interpret each other's intentions, the type of work that we get out of directed graph approaches like David Danks at Carnegie Mellon University, another colleague in the audience that this might come together to help us realize and given domains, lots of general purpose reasoning capability.

Sarah Rajtmajer: Thank you.

Brian David Johnson: Yeah, so I will answer this as an applied futurist and as an engineer. I'm not a computer scientist, I'm an engineer. I like to tell people scientists and mathematicians understand the music of the universe and engineers just build stuff. I build stuff, that's my job. When it comes to this my question, when it comes to general intelligence, I ask people what are you optimizing for, like why, what are you building. I think what Bill was saying one of the things that we're seeing is one of the things I do think will be coming is somewhat limited cognitions where you can actually have these systems, so that you can deploy them to go do a task, so they can learn, they can see triggers, they can then make decisions, then do another step.

I think this is the thing that becomes really interesting and especially when you're looking at the weaponization of artificial intelligence, and let's be clear. All these systems that we're talking about today, everything will be weaponized. We have to know that going into it. If we have time, I'll tell you a story about how I scared a whole bunch of engineers at the consumer electronics show and I told them

they had made one of the best surveillance weapons I had ever seen. They went, "Excuse me?" Then I explained to them how they had and they're like oh and then the press person put themselves between me and the engineers.

For me, I think when it comes to general intelligence or smarter than human AI, I think that's very, very far out and for me oftentimes, I'm just saying why are you building this, like what do you want to do with it. Ultimately, these are tools. Again, this is not a philosophical conversation. We're not trying to replicate humanity, we're just trying to make stuff, and that's one of my key triggers over the next ten years is looking for these systems that have a very limited cognition so that they can take in this information over parameters, then make decisions and then inform other parts of the system to be able to go and do it. I do think that, I'm just coming.

Kevin Chan: I would say I'm a bit more skeptical on the outlook on general intelligence. I mean just from the examples that Patrick was talking about in terms of the models learn from what the information that you give it and the notion of innovation and creating other concepts. From an engineer's perspective, that seems to be very difficult. I think one example is the modeling DNA or the human genome. I mean that has been done, but now looking at how would you model the human brain and going to some of the work in network science, the current state of what they've been able to model in the human brain and then assuming you could model the cognizant behind it is that they've been able to model like the C. elegans

which is like this ridiculously small bacteria, and that's as far as they've gotten.

Kevin Chan: I know that computing and a lot of these so that facilities have helped us do a lot more with more data. It seems to be that there has to be orders of magnitudes of increase of understanding and capability to get there.

Patrick McDaniel: Yeah. To start, I'll follow on something Bill said at the beginning is that these systems are not well understood today. We don't have the mathematical machinery to really understand, for example, deep learning. Deep learning itself, Bill mentioned something called explainability and explainability is basically if your AI says look at a picture and says that's a bird. We don't have any way of knowing why it thinks it's a bird, right? There's some recent work that's starting to get into space. In fact, some of my collaborators at Stanford are making some good progress in this space, but we are absolutely in our infancy with the understanding of the underlying mathematics.

I think one of the things we should be very careful for is not to equate what we're doing and saying. Neural network in our brain, the neurons in our brain are infinitely more complicated than the neurons that are actually in a deep learning system. There are much more complex relationships and physiology and connectivity that are happening inside your brain. To try to equate those two things, we shouldn't do that because they really are different things. Bill, I think we're in vehement agreement in the definition of what we mean by cognition. I think basically the kinds of cognition that I'm thinking are generalized

intelligence, the ability to interact in a world, to learn from a world in the same way that as you mentioned the child would without human intervention.

We could get into really philosophic questions about self-awareness and that's centuries and centuries away. I do agree that there are systems under certain definitions that exist today in limited, what we refer to as constrained domains, but I think getting to the generalized intelligence that you can truly be autonomous in an unconstrained way is a long way off for science and for mathematical reasons and also for other technical reasons.

William Casebeer: Alasdair MacIntyre has a really nice book called Dependent Rational Animals where he makes the point that along the lines of the no true Scotsman fallacy, that there's a sense in which none of us are truly autonomous, right? I'm just wondering if we're setting our standards so high such that people can't meet them either. I'm not sure that I demonstrate cognition on these theories of cognition that are on offer right now, at least I don't feel like I am. I don't know, something to think about.

Sarah Rajtmajer: Maybe our very last panel of the day on autonomy and humanity will touch upon that, perhaps we'll have to see. We only have a few minutes unfortunately, so rather than me ask the questions, let's see if we . . . I see already a hand, so yes.

Audience: I think we have to be careful about bringing too much anthropomorphism to the discussion, especially about AGI. Humans tend to impute, intentionality, all kinds of characteristics to machines that have complex

behavior. We give our cars names. There's lots of psychological literature on that and it colors the way we think about it. AI will not be born, it will not go to school, will not have childhood friends or get bruised knees or solve playground dilemmas, okay. Its experiences are very different. It will not in any way resemble human intelligence in that respect. It will have its own experiences and I think a more productive way to think about it is not a generality, but most as idiot savants.

Nobody here is general intelligence. We have a little common sense if we're lucky, okay, and we can reason about physics in the world a little, but that's about as general and about human relations. This is the big thing that bothers me is that because they're not social, we're going to have a hard time relating and this is a key question. How do we have those productive team discussions and be teammates with AIs that have a different world?

This is our first alien intelligence by the way, this is first encounter. We know it's coming, it's going to be very different, how do we recognize it, how do we make it a productive part of society. I mean I think that resonates with some of you. Others are going to say, "Wait a minute, what about biology?" I don't think you need biology in this. You don't need wetware. My comment, thank you.

Sarah Rajtmajer: Thank you.

Audience: Perhaps everybody in the notion of the rugged individual and the humanist bent as a working mother, I'm not an individual and I'm not autonomous, so here we were just starting to get to questions of whether humans are

autonomous, whether machines are autonomous, but really we function as families, tribes, organizations, and networks. We've seen this also in testimony to congress by people such as chief scientists of the Air Force noting that they do not want machines to be autonomous, they want machines to function within their broader network.

With that then, a lot of the discussion that we had this morning was about making each individual machine able to interact on its own, but I would welcome your comments about how we get machines to function together for collective performance and experience.

Brian David Johnson: There's some work of a colleague of mine at Intel and I'll just keep it really short because I know we're low on time. She painted a world of autonomous vehicles so on the street and the way that she built the architecture, she called it gossiping cars and I loved this, right? As an architect, I was like, "Oh, that makes perfect sense," right? My car can talk to the other car and go, "Oh this guy's jet lag, you should stay away from him and do this," that you create these networks that gossip.

Brian David Johnson: I liked actually just that idea of it because it does tie into what you were saying that these are a broader networking to me as creating autonomous cars that yeah, they're autonomous but they're actually talking to a much, much broader system and creating a language to do that, I think it'd be great so I completely agree of what you're saying.

Patrick McDaniel: I'll follow that on and be my usual negative self. The problem with those systems is you get into a whole world of trust assumptions between

the different systems. There's some great example, so swarm computing was a huge deal maybe fifteen years ago and one of the challenges of swarm community, that's where you have lots of very simple kinds of programs all working in coordination, and the problem with that is that if you have one of them that goes bad, either fails in some way or becomes malicious, they can convince everyone else of an entirely different reality and bring the whole system down.

When you get into collective computing, you have to start getting into discussions about how do you trust what you hear from the other ones, when you're hearing conflicting information, how do you de-conflict that information. There's a whole lot of additional technical challenges that appear when you start getting into distributed independent but cooperating computing.

William Casebeer: Welcome to the world of fake news.

Patrick McDaniel: Yeah.

Sarah Rajtmajer: Okay. Well, I would like to thank our speakers, our panelists for their insights, and thank you for joining us here on this panel and let's move forward to our next set of panels. Thank you.